The Role of Linguistics in Probing Task Design

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

Dissertation

zur Erlangung des akademischen Grades Dr. rer. nat.

vorgelegt von
Ilia Kuznetsov
geboren in Moskau, Russland

Tag der Einreichung:	31. August 2021
Tag der Disputation:	20. Oktober 2021
Referenten:	Prof. Dr. Iryna Gurevych, Darmstadt, Germany Prof. Dr. Martha Palmer, Boulder, USA Prof. Dr. Joakim Nivre, Uppsala, Sweden

31. August 2021
D17

Ehrenwörtliche Erklärung[1]

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades "Dr. rer. nat." mit dem Titel "The Role of Linguistics in Probing Task Design" selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 31. August 2021

Ilia Kuznetsov

[1] Gemäß §9 Abs. 1 der Promotionsordnung der TU Darmstadt

Wissenschaftlicher Werdegang des Verfassers[2]

10/05 – 07/10	Diplom, Theoretische und Angewandte Linguistik, Lomonossow-Universität (Moskau, Russland)
11/12 – 08/16	Aspirantur, Kandidat der Wissenschaften; Mathematische und Computerlinguistik, Higher School of Economics (Moskau, Russland)
10/15 – heute	Doktorand am Fachgebiet Ubiquitous Knowledge Processing (UKP-Lab), Technische Universität Darmstadt

[2] Gemäß §20 Abs. 3 der Promotionsordnung der TU Darmstadt

Anmerkungen zum Umgang mit Forschungsdaten

Gemäß der “Leitlinien zum Umgang mit Forschungsdaten” der Deutschen Forschungsgemeinschaft[3] wurden alle im Zusammenhang mit dieser Dissertation entstandenen Forschungsdaten langfristig archiviert und sofern möglich öffentlich zugänglich gemacht. Folgende Forschungsdaten wurden frei verfügbar gemacht:

- Software
  - Die für die in Abschnitt 2 beschriebenen Experimente notwendige Software steht unter der Apache-Lizenz 2.0 unter https://github.com/UKPLab/coling2018-wcs zur Verfügung.
  - Die für die in Abschnitt 3 beschriebenen Experimente notwendige Software steht unter der Apache-Lizenz 2.0 unter https://github.com/UKPLab/emnlp2020-formalism-probing zur Verfügung.
  - Die für die zusätzlichen Experimente im Abschnitt 3 sowie Abschnitt 4 notwendige Software steht unter der Apache-Lizenz 2.0 unter https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2785 zur Verfügung.
- Forschungsergebnisse
  - Alle im Zusammenhang mit dieser Dissertation stehenden Publikationen sind in der ACL Anthology (http://aclanthology.info/) verfügbar.
  - Alle Forschungsergebnisse sind zudem auch in dieser Dissertation selbst dokumentiert, die von der Universitäts- und Landesbibliothek Darmstadt zur Verfügung gestellt wird.

[3] http://dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf

Abstract

Over the past decades natural language processing has evolved from a niche research area into a fast-paced and multi-faceted discipline that attracts thousands of contributions from academia and industry and feeds into real-world applications. Despite the recent successes, natural language processing models still struggle to generalize across domains, suffer from biases and lack transparency. Aiming to get a better understanding of how and why modern NLP systems make their predictions for complex end tasks, a line of research in probing attempts to interpret the behavior of NLP models using basic probing tasks. Linguistic corpora are a natural source of such tasks, and linguistic phenomena like part of speech, syntax and role semantics are often used in probing studies.

The goal of probing is to find out what information can be easily extracted from a pre-trained NLP model or representation. To ensure that the information is extracted from the NLP model and not learned during the probing study itself, probing models are kept as simple and transparent as possible, exposing and augmenting conceptual inconsistencies between NLP models and linguistic resources. In this thesis we investigate how linguistic conceptualization can affect probing models, setups and results.

In Chapter 2 we investigate the gap between the targets of classical type-level word embedding models like word2vec, and the items of lexical resources and similarity benchmarks. We show that the lack of conceptual alignment between word embedding vocabularies and lexical resources penalizes the word embedding models in both benchmark-based and our novel resource-based evaluation scenario. We demonstrate that simple preprocessing techniques like lemmatization and POS tagging can partially mitigate the issue, leading to a better match between word embeddings and lexicons.

Linguistics often has more than one way of describing a certain phenomenon. In Chapter 3 we conduct an extensive study of the effects of lingustic formalism on probing modern pre-trained contextualized encoders like BERT. We use role semantics as an excellent example of a data-rich multi-framework phenomenon. We show that the choice of linguistic formalism can affect the results of probing studies, and deliver additional insights on the impact of dataset size, domain, and task architecture on probing.

Apart from mere labeling choices, linguistic theories might differ in the very way of conceptualizing the task. Whereas mainstream NLP has treated semantic roles as a categorical phenomenon, an alternative, prominence-based view opens new opportunities for probing. In Chapter 4 we investigate prominence-based probing models for role semantics, incl. semantic proto-roles and our novel regression-based role probe. Our results indicate that pre-trained language models like BERT might encode argument prominence. Finally, we propose an operationalization of thematic role hierarchy - a widely used linguistic tool to describe syntactic behavior of verbs, and show that thematic role hierarchies can be extracted from text corpora and transfer cross-lingually.

The results of our work demonstrate the importance of linguistic conceptualization for probing studies, and highlight the dangers and the opportunities associated with using linguistics as a meta-langauge for NLP model interpretation.

Acknowledgments

As I – somewhat randomly – decided to study linguistics in 2005, I would have never imagined to find myself here, now, finishing an academic text this long dedicated to this topic. There are many people who accompanied me on this path and to whom I owe thanks.

I would like to thank Prof. Dr. Iryna Gurevych for her feedback, guidance and patience over the years, that allowed me to find and develop myself as a researcher, try things out, and keep a broad horizon. I would like to thank FAZIT Stiftung for their funding and support, without which I would have never learned as much.

I would like to thank my past and current colleagues at UKP Lab for their insights, support and hallway chats throughout the years. I wish I joined the lunch more often :)

I am deeply grateful to my professors at the the Moscow State University's Theoretical and Applied Linguistics Department, who over the period of my studies in 2005-2010 gave me a truly broad and multi-faceted education in the very beginning of my academic path. Many of them have left us since then, and I consider it great luck to have learned from those extraordinary researchers.

I would like to thank my former colleagues at the Higher School of Economics in Moscow, and in particular my Candidate of Sciences dissertation supervisor and former colleague Anastasia Bonch-Osmolovskaya for her guidance, advice and having a much better idea about me as a researcher than I myself had at that moment.

Finally – and most importantly – I would like to thank my grandmother Anna for giving me a taste for a good debate and critical thinking, and my parents Oleg and Marina for their never-ending support and encouragement on my path, even when it was hard. Thank you.

Contents

Chapter 1

Introduction

1.1 NLP and Linguistics

Over the past decades natural language processing has evolved from a niche research area into a fast-paced, technically involved, multi-faceted discipline that attracts thousands of contributions from academia and industry and feeds into real-world applications that define the way we interact with information in the digital age.

The general goal of natural language processing is to provide a seamless interface between ambiguous, underspecified, evolving human language and computers. As language is the primary means of human communication, teaching machines to speak and understand is a fundamental milestone for AI with transformative implications for human-computer interaction and processing and handling of human knowledge. Despite the progress of the past decades, this goal is far from being reached: while impressive results have been achieved for simple tasks in-domain, modern natural language processing models still struggle to generalize across domains (Beltagy et al., 2019), are vulnerable to adversarial attacks (Nie et al., 2020), resort to low-level heuristics leading to biases (Tan and Celis, 2019) and often fail to perform simplest tasks that only require basic linguistic competence (Ribeiro et al., 2020). While some of those issues can be partially addressed on-site by acquiring more labeled data, refining the training protocols and incorporating common-sense knowledge, the fundamental understanding of the mechanisms that lead to improvements is often lacking, as is the guarantee that a said model truly reflects the way human language works and will thereby transfer well to new use cases and domains.

Language is one of our key abilities and most valuable assets. Despite the abundance of natural languages, there exist numerous regularities in how languages represent meaning, express grammatical information and construct utterances. Linguistics studies language as a general system, aiming to discover how language is acquired and used, how it evolves over time, how it is structured and which universal principles guide the use of language as a whole, independent of the particular language instance. Theoretical linguistics focuses on the fundamental principles behind human language as a phenomenon, while computational linguistics aims to explore the use of language and verify linguistic hypotheses using machine-assisted methods.

Modern natural language processing originates from work in computational linguistics and information retrieval; however, the application-driven incentive of NLP has steered it in a more practical direction, and although natural language processing and computational linguistics coexist in the research landscape and often share publication venues, study outcomes and research staff, their goals and research methodology diverge. Linguistic discovery – happening at a lower pace than the discovery in the computational domain – has not been the main driving force behind the recent NLP successes, and yet linguistics is omnipresent in modern NLP in the form of objectives and evaluation criteria, resources and auxiliary signals, and principled analytical frameworks. As mainstream NLP gradually moves towards strong pre-trained neural language models and end-to-end processing setups, our work aims to offer a timely reflection on the role of linguistics in this process.

1.2 From Rules to Probing

To put our work in a broader context, we start with a short overview of the major milestones in NLP system development and evaluation methodology.

Machine learning and benchmarks. Since its start, natural language processing has co-evolved with the rest of the informational and computational infrastructure. Early NLP has predominantly been rule- and pattern-based: the rules had to be manually constructed by experts and often delivered acceptable performance for closed-domain, simple application scenarios. Rule-based systems have an additional advantage of being interpretable: the automatic predictions can be traced back to the symbolic rules contributing to these predictions; these rules can be used as justification (in case of correct predictions) or as guidance for adjustment (in case of incorrect predictions). Creating and maintaining the rule base, however, is an extremely time-consuming task that requires expertise, and the progress in machine learning and availability of annotated corpora have given rise to simple machine-learning based natural language processing systems. Methodologically, the availability of shared annotated resources has made it possible to apply the standard information retrieval evaluation machinery and metrics to natural language problems: NLP approaches could now be compared in terms of their benchmark scores, and the better-performing approach would be considered state of the art. Benchmarking and performance-based evaluation have provided a solid way to track NLP progress and have since become the main driving force of NLP, leaving qualitative evaluation, hypothesis-driven research and error analysis as secondary – yet important – role in steering the field.

Features and pipelines. Early NLP approaches have operated with discrete features often derived from linguistic and surface-level properties of the underlying texts: for example, a rule-based named entity recognition module would utilise keyword gazetteers, capitalisation of the tokens, prepositions and contextual cues like the word to the left and to the right of a potential named entity. Not all of these features were directly accessible from the surface text, motivating the use of NLP pipelines: a set of upstream NLP components would produce annotations on top of the original text – for example, part-of-speech tags or syntactic groupings – which would be used for feature extraction in the downstream modules. This made the prediction process straightforward and clearly separable, however, the one-way nature of the NLP pipelines posed a severe limitation on the resulting systems: the prediction quality of an NLP component is never perfect, and the errors made in earlier steps of the pipeline propagate into the latter components without any opportunity for the NLP system to correct the upstream predictions post-factum. This has motivated the development of joint models that would learn to simultaneously make predictions for multiple steps of the pipeline, and have consistently outperformed their step-wise counterparts – at the price of additional restrictions on the task and data setup and reduced interpretability of the resulting models.

Continuous representations. Despite being easy to interpret, discrete features have important drawbacks. They might become inefficient in terms of storage and computation, resulting in sparse inputs which are not well-suited for the majority of machine learning algorithms: for example, introducing a feature corresponding to the lemma of a given word would require creating a binary vector representation with dimensionality of the vocabulary size, out of which only one value would be used at a time. Another drawback of discrete features is their inability to model similarities between feature values: in a discrete feature space a "man" is as similar to a "woman" as it is to a "car". This severely limits the generalization capability of the models building upon such representations, and in many cases it would be useful to represent the input word space in a more structured way. One solution to this issue is clustering, however, a more generalizable approach is to map discrete features into a lower-dimensional space which models the similarities between these features: this technique has been widely implemented and led to – often dramatic – increases in end-system performance, at the cost of the further interpretability loss: a dense vector representation of a feature cannot be directly mapped back to its human-readable and interpretable value.

Transfer learning. The re-emergence of neural network-based approaches has started a new era in natural language processing: the ability of neural networks to model non-linearities in a computationally efficient way, the know-how inherited from neighbouring fields like computer vision and the development of architectures specifically tailored for language processing have allowed for more flexible modelling of language and boosted the performance to the point when simply trading an SVM for a multi-layer perceptron would lead to a substantial performance increase. However, the arguably more important development in NLP of the past decade responsible for most of the field's recent successes is transfer learning, in which the internal model knowledge learned during pre-training on a related task is re-used to improve predictions on the target task. The growth of the Internet and the abundance of easily accessible plain text have made words (more specifically, word types) the primary target for transfer pre-training: a large text collection coupled with an unsupervised language modeling objective could be used to produce word embeddings that would exhibit meaningful properties, and – given an efficient implementation like word2vec (Mikolov et al., 2013b) – would not require too many resources to compute. A multitude of word embedding approaches appeared, which would often be coupled with one or several pre-trained vector space model instances ready to be used in downstream NLP systems: this provided the researchers with a shared representation that would be tailored to a particular task at hand. While word type embeddings have proven useful on most occasions, the relationship between the positions in the resulting vector spaces and the linguistic phenomena they corresponded to remained opaque. A multitude of studies have aimed to investigate these correspondences (Hill et al., 2015; Gerz et al., 2016) and to make word embeddings more "linguistically aware" either by modifying the pre-training objective (Ebert et al., 2016) or by retrofitting the pre-trained space using a linguistic signal (Faruqui et al., 2015).

Deep models. For a long time the representation offered by static pre-trained word embeddings has been useful and yet too generic to be directly utilized for solving downstream tasks. As a result it had to be coupled with a task-specific deep architecture, which would use the pre-trained static word embeddings as a source and combine them with embeddings for other features that would be learned during training for the target task. The resulting systems again outperformed shallow models of the past, however, the depth of the architecture has rendered the analysis of the source static word embeddings less viable: whatever linguistic properties these pre-trained vector spaces would represent could be overwritten by the downstream layers of the deep network, and a large bulk of the performance gain would originate from the task-specific model itself.

Strong encoders. In the meantime pre-training word type representations with unsupervised language modeling objectives has progressed, resulting in the introduction of strong, contextualized, subword-level encoder models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). Due to their representational capacity and a careful choice of the pre-training objective and regime, the new contextualized representations are able to encode much more information than their static word-based counterparts; moreover, the resulting representations are often multi-layered – providing different levels of abstraction over the same sequence that the end architecture could utilize – and dynamic – implicitly disambiguating the input tokens in a context-dependent manner. Modern contextualized pre-trained encoders encode so much general-purpose information that using a deep downstream neural architecture often becomes unnecessary: the state of the art on many natural language processing tasks is held by a strong pre-trained encoder followed by a simple task-specific classifier. Shifting a large bulk of the model's capacity to a pre-trained representation has several advantages: if the pre-trained encoder is not updated during training, the task-speficic training and prediction are computationally inexpensive; it allows for easier domain adaptation of NLP systems as long as a sufficient amount of raw textual data in a new domain exist; finally – and most importantly in the context of this thesis – it provides researchers with a shared, re-usable representation.

Probing. As with static embeddings, most encoder models are accompanied by a ready-to-use pre-trained model instance, which gives researchers a wide common ground and renders the analysis of the pre-trained representations relevant as never before. Due to the complexity and depth of the resulting representations, a direct analysis of the models' predictions is challenging. The emerging area of probing aims to assess the linguistic capabilities and properties of the strong pre-trained encoders using probing tasks – simple, restricted natural language processing objectives designed to measure how and if a pre-trained model represents certain basic phenomena. Due to its wide applicability and theoretical well-formedness, linguistics becomes a convenient common point of reference for this analysis. Recent studies demonstrate that pre-trained models indeed rediscover many important linguistic relations and learn linguistically meaningful abstractions solely based on unlabeled text.

As simplified and general as it might be, the outline above demonstrates that over the past decades natural language processing has progressed from simple rule-based systems to intricate task-specific neural architectures to strong pre-trained encoder-based models, gradually trading interpretability and transparency for performance and practical applicability. The emergence of strong pre-trained encoders is a transformative step in the development of natural language processing, not only due to the higher benchmarking scores it allows, but, more importantly, due to providing NLP practitioners with a strong common representation that is more powerful than discrete pipelines of the past, is more transferable than task-specific NLP models, and displays the ability to represent certain linguistic regularities, albeit in a non-conventional way.

1.3 This work

In the context of the fields' transition towards end-to-end systems powered by strong pre-trained encoders, our work takes a step back and examines how linguistic conceptualisation structurally affects the design, evaluation and interpretation of the modern pre-trained NLP models. Linguistics as a discipline operates at a lower pace than natural language processing and machine learning, and from an NLP perspective linguistics is a rather static body of knowledge to be drawn from. Despite that, linguistics has major influence on the development of the field, both explicitly – by providing resources, theory operationalisations, benchmarks and abstraction layers, and implicitly – by framing the tasks and defining the objectives for the NLP models to reach. In this work we investigate how NLP and linguistics interact in benchmarking (RQ1), probing (RQ2) and task design (RQ3).

To keep our analysis tractable, we focus our studies on a particular linguistic domain – predicate and role semantics. Predication is the minimal logical unit of language that can be assigned a truth value and positioned on the time and evidentiality scale, and if we aim to develop machines able to understand and reason over texts, such machines would inevitably need to capture predicate semantics, at least implicitly. One family of predicate-semantic formalisms – role semantics – has a long history of research in linguistics, is well operationalised in NLP in form of semantic role labeling (SRL), covers both grammatical (syntax) and semantic (lexical semantics) phenomena and is positioned on the border between grammar – which modern pre-trained encoders seem to capture well – and high-level semantics – which still remains challenging. This makes role semantics an attractive research objective on its own, and an ideal target for our investigations. Our findings contribute to the lexical-semantic evaluation of predicate similarity (RQ1), encoding of role-semantic phenomena in pre-trained language models (RQ2), and modeling of syntactic and semantic prominence in NLP (RQ3).

1.3.1 Research questions

This thesis addresses the following research questions.

RQ1 Conceptual gap between word embeddings and lexicons Pre-trained static word embeddings have had a major impact on almost every research area in NLP. A standard way to compare and evaluate static word embeddings is similarity benchmarks: the similarities between word encodings in a pretrained vector space model are compared to the similarity scores assigned to word pairs by humans in terms of correlation. Higher correlation coefficients signal that a pre-trained model captures the lexical similarity well, while lower correlation points at potential issues. However, there exists a major conceptual discrepancy between the two entities under comparison: classic word embedding approaches operate with word types, while standard human-assessed benchmarks implicitly operate with disambiguated, lemmatized lexical units. **What are the effects of the conceptual gap between static word embeddings and similarity benchmark targets?** We investigate this question in Chapter 2.

RQ2 Effect of linguistic formalism on probing Linguistics often has multiple ways to represent and describe a certain phenomenon: for example, there exist several major POS tagsets, several widely used dependency formalisms and several role-semantic frameworks that differ in terms of granularity, internal organization, underlying assumptions and scope. Recent work in probing for deep contextualized pre-trained encoders reveals that models like BERT implicitly learn a great amount of linguistically relevant information during unsupervised pre-training, and that the order in which information is processed in the layer structure of such models aligns with the structure of the traditional NLP pipeline (Tenney et al., 2019a). However, while claiming that, for example, BERT discovers syntax or POS categories, a probing study in fact only shows that BERT's internal representations can be mapped to a specific linguistic *formalism*, e.g. English Universal Dependencies syntax and universal POS categories. **Can the choice of linguistic formalism have an effect on the probing results?** We address this question in Chapter 3.

RQ3 Role labeling and Prominence Most work on predicate and role semantics in natural language processing has focused on locally assigning roles to the semantic arguments of predicates based on syntactic and lexical information. Global interactions between semantic roles on a sentence level have been a major subject of study in linguistics and motivate many decisions in the semantic role theories; however, in NLP these interactions have so far been modeled as a purely formal constraint, e.g. two arguments of the same predicate cannot be assigned the same semantic role. Prominence and thematic hierarchy are two linguistic concepts used to account for role interactions and syntactic realization. Despite the years of research this concept has not yet been operationalized for use in NLP and evaluated on corpus material. **How can we operationalize prominence and thematic hierarchy for use in NLP, and what can we learn from prominence-based probing of the modern pre-trained encoders?** We investigate these questions in Chapter 4.

1.4 Contributions

RQ1 Conceptual gap between word embeddings and lexicons

- We hypothesize that the conceptual gap between the word embedding vocabularies and the entries in similarity benchmarks and lexical resources can lead to decreased benchmarking performance.

- We investigate this hypothesis using standard similarity benchmarks and a novel resource-based word class suggestion scenario which targets some of the drawbacks of the existing resource-based evaluation methods. We apply the novel evaluation scenario to two resources: predicate-centered VerbNet (Schuler, 2005) and general-purpose WordNet (Miller, 1995).

- We show that lemmatizing and POS-disambiguating the word embedding targets indeed leads to improved benchmark and resource-based performance, with the effect especially pronounced for English verbs.

RQ2 Effect of linguistic formalism on probing

- We hypothesize that the differences in linguistic formalism might affect the results of probing deep pre-trained contextualized encoders. To investigate, we compare probing results for three role-semantic formalisms – PropBank (Palmer et al., 2005a), VerbNet (Schuler, 2005) and FrameNet (Baker et al., 1998) – using rich parallel annotated data. We validate our findings by experimenting with a multilingual pre-trained model on two languages: English and German.

- We refine the recently proposed edge and layer probing framework (Tenney et al., 2019b), and expand it by introducing anchor probing tasks that allow qualitative insights into the representations learned by strong pre-trained encoders and allow the incorporation of expert-designed feature sets from early NLP systems into the analysis.

- We confirm and extend previous findings on the sequential processing order and abstraction strategy within pre-trained contextualized encoders in a multilingual setting.

- We show that the choice of formalism affects the layer probing results in a linguistically meaningful manner. By allowing the probing model to learn to extract separate representations for predicate and argument tokens in the role probing scenario we find that while the predicate-agnostic PropBank-style probe focuses on the higher layers of the pre-trained model associated with syntax, Verbnet and FrameNet-style role labeling probes implicitly model predicate semantics using the lower layers. Anchor task analysis qualitatively mirrors the feature sets used for predicate disambiguation and role labeling in the early SRL systems.

- In a set of additional experiments, we investigate the effects of training data size, dataset and task granularity on layer probing results. Our results provide important insights into the aspects of the layer probing relevant to future cross-formalism probing studies.

RQ3 Role labeling and Prominence

- We refine the previous studies of the encoding of semantic proto-role properties in contextualized language models and show that some of these properties can indeed be extracted from the pre-trained BERT, contrary to the existing reports.

- We suggest a novel regression-based role labeling probe to investigate whether pre-trained contextualized encoders capture semantic prominence relations. We show that role labelling can indeed be cast as a regression task; our results provide new evidence for layer encoding properties of BERT models w.r.t. predicate semantics. Our evidence suggests that pre-trained BERT models implicitly capture the notion of prominence and show a preference for viable semantic role hierarchies.

- We suggest a framework for inducing thematic hierarchies from syntactically and role-semantically annotated corpora, as well as evaluation criteria and a range of methods for thematic hierarchy induction. We evaluate it on English and German data and demonstrate that thematic hierarchies can be induced from small amounts of training data and apply cross-lingually.

1.5 Publication record

Parts of this thesis have been previously published and presented at international peer-reviewed conferences. Some of these published works serve as foundation for the text presented here and are reused verbatim throughout the thesis; others have contributed to the studies reported here indirectly.

- In *Assessing SRL Frameworks with Automatic Training Data Expansion* (Hartmann et al., 2017a) together with the colleagues from the UKP Lab and the Heidelberg University we have experimented with augmenting the training data for semantic role labeling using annotation projection techniques. The paper introduces the SR3de corpus created by our colleagues and used in this thesis.

- In *Out-of-domain FrameNet Semantic Role Labeling* (Hartmann et al., 2017b) we helped investigate the out-of-domain performance of FrameNet-style semantic role labelers, and found that predicate disambiguation is a major bottleneck in the frame-semantic parsing pipeline. The paper proposes a simple method based on static word embeddings that partially alleviates the issue. While this work is not used in the thesis directly, it served as inspiration for our later role probing study; the high importance of predicate semantics to VerbNet and FrameNet-style role labeling is echoed by our probing results in Chapter 2.

- In *From Text to Lexicon: Bridging the Gap between Word Embeddings and Lexical Resources* (Kuznetsov and Gurevych, 2018a) we investigate the conceptual discrepancies between static word embedding vocabularies and show how these discrepancies affect benchmarking results. This publication forms the basis of Chapter 2 and its passages are quoted verbatim.

- In *Corpus-driven Thematic Hierarchy Induction* (Kuznetsov and Gurevych, 2018b) we propose a framework for analyzing and inducing thematic hierarchies from corpus data and evaluate it using VerbNet-style role inventories on English and German. Our results are used as part of Chapter 4 and quoted verbatim.

- In the collaboration *LINSPECTOR: Multilingual Probing Tasks for Word Representations* (Şahin et al., 2020) we helped create a large-scale multilingual probing suite for static word embeddings. Similar to the layer probing approach used in this thesis, LINSPECTOR features a layer-wise analysis of representations with respect to a range of isolated linguistic phenomena. Our contribution to this work, however, was on the task design and multilinguality side, and no parts of LINSPECTOR text, code or resources have been reused in this thesis.

- Finally, in *A Matter of Framing: The Impact of Linguistic Formalism on Probing Results* (Kuznetsov and Gurevych, 2020) we investigated the effects a linguistic formalism can have on the probing measurements. We show that linguistic formalism can affect probing results and our insights warn against overgeneralising claims made while probing with linguistic material. This source forms the core of Chapters 3 and largely contributes to Chapter 4, both of which quote it verbatim.

Apart from the research directly related to the thesis, we have participated in several smaller projects dedicated to computational analysis of scientific literature and NLP for peer reviewing, two of which have led to additional publications. In *EELECTION at SemEval-2017 Task 10: Ensemble of nEural Learners for kEyphrase ClassificaTION* (Eger et al., 2017) we implemented a keyphrase classification approach for scientific terms. *Does My Rebuttal Matter? Insights from a Major NLP Conference* (Gao et al., 2019) investigates the effects of conformity bias on peer reviewing results using the data from the ACL-2018 conference, and proposes new NLP-based approaches to model the effectiveness of author rebuttal. These works do not contribute to the material presented below.

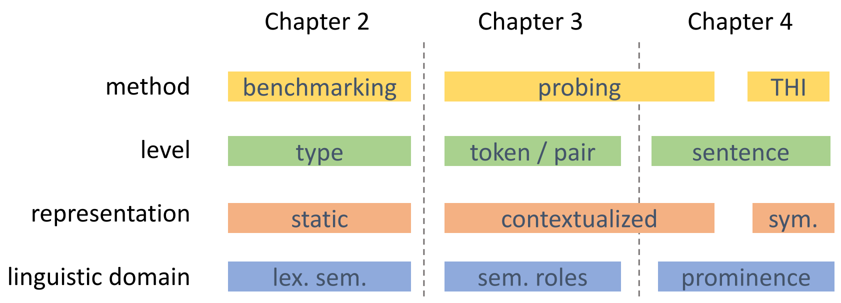


Figure 1.1: Thesis structure

1.6 Thesis overview

Our work covers a lot of ground, from static word embeddings and basic similarity benchmarks to modern multilingual contextualized encoders and state-of-the-art probing methodology, and from basic linguistic phenomena like part of speech to the high-level notions of semantic roles and role prominence that reside at the border of syntax and semantics. Below we review the structure of this thesis, and Figure 1.1 serves as a simplified reference to the research objects, methodologies and linguistic phenomena used throughout this work.

From the **methodological** perspective, probing makes up the core of this work, with Chapter 3 and parts of Chapter 4 directly using and building upon the state-of-the art layer probing methodology to investigate how linguistic information is represented by the pre-trained encoders. Chapter 2 builds upon similarity and resource-level benchmarking, which, as we argue, can also be seen as a variation of probing. Finally, we propose a new non-neural method for thematic hierarchy induction from syntactic data in Chapter 4.

Based on the target **level** and granularity, the thesis progresses from static type-level tasks well suited for modeling lexical semantics in Chapter 2, to contextualized token- and token pair-level tasks that focus on grammar and disambiguation in context in Chapter 3; finally, in Chapter 4 we move to sentence level to investigate the interdependencies between semantic roles of the same predicate.

The choice of the modeling target motivates our choice of **representations**: for type-level tasks in Chapter 2 we use static word embedding models (in particular, different configurations of word2vec (Mikolov et al., 2013b); for token and token-pair level tasks in Chapters 3 and 4 we build upon the contextualized, deep, Transformer-based BERT representations (Devlin et al., 2019); finally, the hierarchy induction experiments in Chapter 4 use traditional, symbolic representations of syntax.

Finally, from the **linguistics** perspective, Chapter 2 is dedicated to lexical semantics with a special focus on verbs; Chapter 3 uses semantic role assignment in three major role labeling formalisms to investigate the impact of formalism on layer probing; Chapter 4 investigates multiple approaches to modeling semantic prominence in the context of probing, including semantic proto-role probes (Reisinger et al., 2015a) and our novel regression-based role labeling probes that depart from the NLP tradition of modeling semantic roles as categorical, atomic classes (Gildea and Jurafsky, 2000). We conclude our discussion on prominence with a feasibility analysis of the full thematic hierarchy induction from corpus data in Chapter 4.

Chapter 5 summarizes the thesis and concludes the work with some final remarks and future research suggestions.

Chapter 2

From Text to Lexicon

Despite the recent advances in contextualized encoding of words and sentences, static word embeddings maintain a strong presence both in research and in practical natural language processing due to their simplicity and low resource demands. However, what do the *"words"* targeted by the word embedding models in fact correspond to in a linguistic sense, and do the common similarity- and resource-based evaluation setups align with that notion? In this chapter we investigate the conceptual gap between the entries of word embedding vocabularies and the lexical resources. We find that the conceptual discrepancy between the notions of word type and lexeme have practical consequences for evaluation of static pre-trained embedding models.

2.1 Introduction

The training objective in the unsupervised static word embedding setup is to induce vector representations for *targets* based on their *contexts* encountered in an unlabeled, plain text reference corpus. The goal is to encode targets so that the targets appearing in similar contexts are close in the resulting *vector space model* (VSM). We further refer to the set of targets in a given VSM as the *vocabulary* of this VSM. For example, the target *"cat"* might appear in contexts "*A ... sleeps on a mat*" and *"Don't forget to feed the ..."*. The vector associated with this target by the VSM will be close to the vectors of other house pets in the VSM's vocabulary and – ideally – distant from the vectors of country capitals, chemical compound names, etc.

Static word embeddings have been responsible for a major leap in natural language processing performance due to their broad coverage and ability to implicitly encode lexical relations. Seemingly overshadowed by the recent progress in dynamic word and sentence embedding methods in cutting-edge NLP, static word embeddings remain the representation method of choice for many researchers and NLP practitioners due to their ease of use and low computational and infrastructure requirements. In the first half of 2020 alone, static word representations like GloVe (Pennington et al., 2014) have been used in a variety of tasks and systems spanning from dialogue modeling (Ma et al.,

2020) to sentence simplification (Kumar et al., 2020), to automatic analysis of radiology reports in medicine (Ong et al., 2020) and anonymization of clinical notes (Hartman et al., 2020), among many others.

The lasting popularity and the ability to easily pre-train custom static embeddings from arbitrary corpora call for generalizable evaluation methods that would allow assessment and comparison of the resulting vector spaces. While extrinsic task evaluation is often preferable, it does not separate the embedding quality from the end-task model quality and thereby does not contribute insights into generalisability and transferability of the pre-trained models to new tasks. Motivated by this, a long-standing line of research in word embedding evaluation aims to develop methods for task-independent assessment of the pre-trained word embedding models. Two principled approaches for task-independent evaluation of static word embeddings are *similarity benchmarking* in which the distances in the pre-trained vector space model are compared to the word similarity judgements made by human annotators, and *resource-based benchmarking* in which the relationships between word embedding vectors are mapped onto lexical relations codified in a general-purpose lexical resource. Methodologically these benchmarking approaches are direct precursors of the current line of research in probing: given a general, simple linguistic task (e.g. similarity benchmark) and a low-capacity prediction model (e.g. threshold on the raw cosine distance), probing rests on the assumption that the performance on the probing task reflects the internal properties of the pre-trained encoder and the information implicitly captured by it.

2.1.1 Motivation

A probing model should have minimal own capacity to prevent it from learning to predict the phenomenon based on training data and not on the information implicitly encoded during pre-training (Hewitt and Liang, 2019). Many successful approaches to probing and benchmarking are indeed built around directly correlating the raw distances in the pre-trained vector spaces with an external task-specific signal. This, however, allows conceptual inconsistencies in the probing setup to directly project into the benchmarking results, as the model has no capacity to mask and compensate for the discrepancies between pre-trained inputs and human-annotated targets. As an extreme example, imagine that one would evaluate a pre-trained static word-level model like word2vec on a sentence-level semantic textual similarity benchmark (Cer et al., 2017): while a conceptually aligned setup using mean pooling would perform very well (Ranasinghe et al., 2019), assessing sentence similarities solely based on the vectors of their first words would fail due to the mismatch between the encoding target (words) and the evaluation target (sentences). The observed poor performance in this case would not be due to the inability of word-level encoders to capture sentence similarity, but due to the design issues in the probing setup.

As exaggerated as the above example might be, similar effects are present in the traditional static word embedding evaluation settings. Since the further discussion requires precise word-related terminology, we introduce it here. A *token* is a unique word occurrence within context. Tokens that are represented by the same sequence of characters belong to the same *type*. A type signifies one or – in case of morphological ambiguity – several *word forms*. Word forms encode the grammatical information carried by the word and are therefore POS-specific. One of the word forms is considered the *base form* – or *lemma* of the word. All possible word forms of a word represent the *lexeme* of this word. From a semantic perspective, a word is assigned to a minimal sense-bearing *lexical unit* (LU). In general, multi-word and sub-word lexical units are possible, but here we focus on single-word units for simplicity. LUs are the reference point to the semantics of the word and might be used to describe further properties of the words within a specific *lexicon*, e.g. get assigned one or several senses, related to other LUs via lexical relations etc.

With the above hierarchy in mind, let us examine how common static word embedding approaches and evaluation benchmarks define their targets. The applications of word embeddings can be roughly grouped in two categories, *occurrence-based* and *vocabulary-based*: the former aims to classify tokens (e.g. parsing, tagging, coreference resolution), the latter aims to classify lexemes (e.g. word clustering, thesaurus construction, lexicon completion). Lexical resources mostly operate on lexeme or lexical unit level. This includes most of the similarity benchmarks (Finkelstein et al., 2001; Bruni et al., 2014; Hill et al., 2015; Gerz et al., 2016) that implicitly provide similarity scores for lexemes and not for word types. Traditional static word embedding approaches, however, induce representations on the type level. As Figure 2.1 demonstrates, this leads to a conceptual gap between what the word embedding methods learn to represent and what they are evaluated against.

Albeit seemingly subtle, this conceptual discrepancy is problematic for several reasons. Consider the standard scenario where a type-based VSM is evaluated against a lexeme-based benchmark. One of the types contained in the VSM vocabulary corresponds to the base form (lemma), and the vector for the base word form is used for evaluation. This particular form, however, is selected based on *grammatical* considerations and, as we later demonstrate, is neither the most frequent nor the most representative in terms of contexts, nor the most unambiguous. As a result, (1) the contexts for a given lexical unit are under-represented, ignoring other word forms than the base form; (2) in case of ambiguity the same static representation is learned for several different lexemes sharing the same word type. Based on this, one might hypothesise that a better conceptual alignment between the pre-trained embedding models and lexical resources would lead to higher agreement between them and result in more linguistically meaningful pre-trained representations.

Strengthening this assumption, multiple studies have demonstrated that even partially addressing these problems indeed leads to improved performance. For example, Ebert et al. (2016) have introduced the lemma-based LAMB embeddings and have shown that lemmatization of the targets improves the results cross-linguistically on similarity benchmarks and in a WordNet-based evaluation scenario. In our scheme this represents a step up in the concep-

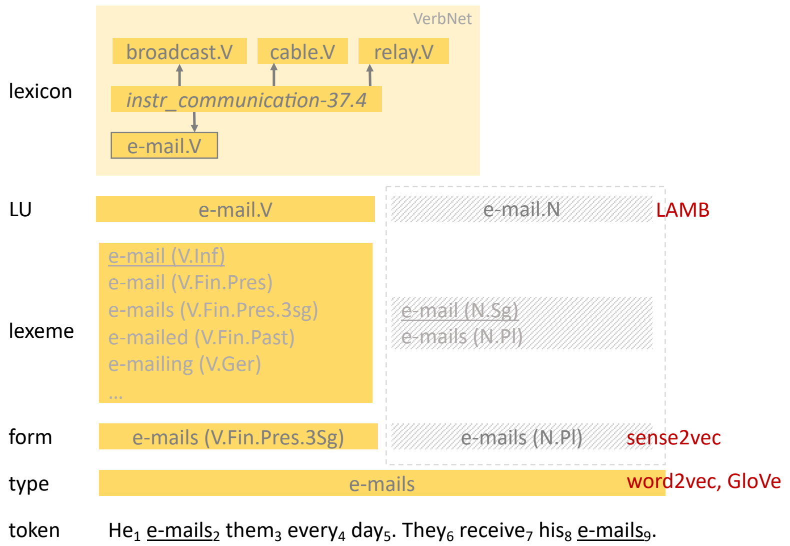


Figure 2.1: Hierarchy of word-related concepts for *emails*, with VerbNet as lexicon and word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), sense2vec (Trask et al., 2015) and LAMB (Ebert et al., 2016) exemplifying static embedding approaches with varying degrees of target disambiguation.

tual hierarchy but still leaves room for ambiguity on the LU level. Trask et al. (2015) experimented with POS-disambiguated targets, and also report improved performance on a variety of tasks.

While these studies support the hypothesis that lemmatization and POS-typing of targets are beneficial for downstream tasks, they do not provide a detailed investigation on why it is the case and do not examine the effects of combining the two preprocessing techniques. Our work aims to close this gap. We evaluate the effects of lemmatization and POS disambiguation on similarity benchmarks in a controlled setup, and further refine our results using a novel resource-based word class suggestion scenario which measures how well a VSM represents VerbNet (Schuler, 2005) and WordNet supersense (Ciaramita and Johnson, 2003) class membership. We find that POS-typing and lemmatization have complementary qualitative and vocabulary-level effects and are best used in combination. We observe that English verb similarity is harder to model and show that using lemmatized and disambiguated embeddings implicitly targets some of the verb-specific issues.

2.1.2 Contributions

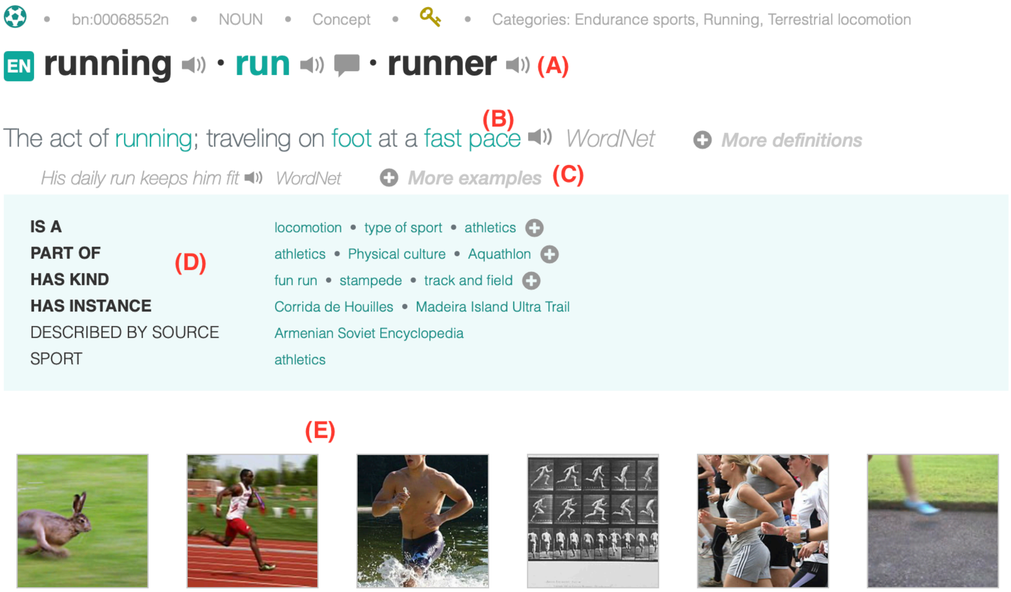
- We suggest using lemmatized *and* POS disambiguated targets as a conceptually plausible alternative to type, word form and lemma-based VSMs;
- We introduce the suggestion-based evaluation scenario applicable to a wide range of lexical resources;
- We show that lemmatization and POS disambiguation improve both

benchmark and resource-based performance by implicitly targeting some of the grammar-level issues.

2.2 Static word representations

2.2.1 Models of Lexical Semantics

The ability to represent and process the meaning of words and sentences is key to the success of natural language processing, as far as any applications beyond purely formal grammatical analysis are concerned. Lexical units are the core semantic building blocks of language and the main subject of study in the field of lexical semantics. Numerous theoretical frameworks and implementations have been proposed in the linguistic literature and subsequently adapted for use in more practical NLP applications. From the natural language processing perspective, there exist two principled approaches to representing the meaning of words and multi-word expressions: lexical resource-based, and distributional.



• bn:00068552n • NOUN • Concept • • Categories: Endurance sports, Running, Terrestrial locomotion

EN **running** • **run** • **runner**

The act of running; traveling on foot at a fast pace WordNet

His daily run keeps him fit WordNet

IS A	locomotion • type of sport • athletics
PART OF	athletics • Physical culture • Aquathlon
HAS KIND	fun run • stampede • track and field
HAS INSTANCE	Corrida de Houilles • Madeira Island Ultra Trail
DESCRIBED BY SOURCE	Armenian Soviet Encyclopedia
SPORT	athletics

Figure 2.2: BabelNet as an example of a modern linked LSR, that given a lexical unit (A) provides its definition (B), usage examples (C), relations to other lexical units and categories (D), as well as visual examples (E).

Resource-based approaches model lexical-semantic knowledge as a finite lexicon in which every lexical unit is assigned a definition, a set of properties, a set of relations to other lexical units (Figure 2.2). Lexicon-based models of semantics have a long history in linguistics and didactics: monolingual dictionaries associating each entry with a textual explanation and basic grammatical information are a staple of lexicography and are widely used in academic research, language teaching and in everyday life. Aside from the traditional definition-based dictionaries like the *Oxford Dictionary* for English and *Duden* for German, this includes recent collaborative efforts like Wiktionary[1] and Urban Dictionary [2].

Despite their popularity, monolingual dictionaries rely on implicit world knowledge and do not exploit structural properties of the lexicon, limiting their utility for automatic natural language processing. This fact has been acknowledged in lexicography early on, and a wide range of structured lexical-semantic resources (LSRs) emerged. The classic example of structured lexical representation is Roget's Thesaurus (Kirkpatrick, 2000): originally intended as a writing aid, it organizes the lexicon into a hierarchy of classes, with each subsequent layer of the hierarchy representing finer semantic distinctions. Another prominent example of a structured lexical-semantic resource is WordNet (Miller, 1995) – an extensive database and network of lexical units grouped into synonym sets (or synsets) and connected with lexical-semantic relations. A structured lexical-semantic resource can be enriched with additional formal information about the semantic and grammatical properties of the entries: for example, VerbNet (Schuler, 2005; Kipper et al., 2006; Schuler et al., 2008) classifies verbs based on their syntactic behavior and associates each class with subcategorization information and selectional restrictions; FrameNet (Baker et al., 1998) follows frame semantics and groups lexical units based on the frames they evoke, each frame accompanied by rich semantic and grammatical metadata. Lexical semantic resources can be interconnected, providing wider coverage and even richer semantic representation; some prominent examples of linked LSRs include BabelNet (Navigli and Ponzetto, 2012), UBY (Gurevych et al., 2012) and SemLink (Bonial et al., 2013; Stowe et al., 2021). Lexical-semantic resources have been widely used in natural language processing both as an additional source of information for the models and as a way to evaluate models in terms of the lexical knowledge they have already acquired from the data or other sources.

Lexical resources provide a rich, theoretically grounded and human-interpretable representation of lexical knowledge. However, the development and use of lexical-semantic resources is associated with a range of drawbacks limiting their utility for natural language processing. Despite their size, lexical-semantic resources often suffer from low coverage, especially when applied to non-standard domains (Hartmann et al., 2017c). Developing an LSR requires substantial expert involvement, limiting their expansion and domain adaptation. Finally, the differences reflected in an LSR, while theoretically sound, might introduce an unnecessary level of complexity for downstream NLP tasks. For example, WordNet is known to be extremely fine-grained, accounting for sense distinctions even human annotators cannot reliably reproduce (Palmer et al., 2007); in response, alternative sense distinction schemes have been successfully developed (Hovy et al., 2006) and alternative granularity levels have been explored

[1] http://www.wiktionary.org

[2] http://www.urbandictionary.com

and integrated into NLP applications (Flekova and Gurevych, 2016).

An alternative approach to representing lexical semantics is based on the distributional hypothesis, which states that the meaning of a lexical unit can be approximated by the contexts this lexical unit appears in (Harris, 1954). As an illustration of this concept, a language user unfamiliar with a word *kohlrabi* and given a small set of contexts like *"How to plant and grow kohlrabi"*, *"Crispy apple and kohlrabi salad"* and *"Roasted kohlrabi with garlic sauce"* can reliably place the new word in their lexicon and infer other properties of the newly discovered root vegetable. Unlike lexical resources, distributional semantic models do not require expert involvement and do not make any structural assumptions about the lexicon. They can be constructed based on unlabeled text collections, largely alleviating the coverage issues associated with LSRs. Due to the availability of unlabeled textual data and efficient implementations, distributional semantic models have become the de-facto standard in natural language processing, and the rest of this chapter will focus on this class of models. Before we proceed, however, it is important to note that distributional models are complementary to lexical-semantic resources: the best end-task performance is often achieved via a clever combination of high-coverage, but poorly structured distributional representations and lower-coverage structured lexical resource data (Faruqui et al., 2015; Bevilacqua and Navigli, 2020).

2.2.2 Sparse models

On an abstract level, building a distributional semantic model requires as input a text collection (corpus), a definition of target and a definition of context. The goal of a distributional semantic model is to create a context-based representation in which similar targets are positioned close in the resulting representation space. For simplicity, in this chapter we focus on single-word targets and omit the discussion on multi-word expressions; however, most of the following argumentation equally applies to targets consisting of more than one word.

A naive approach to distributional modeling is to represent each target word as a number of times it occurs in a certain context. Since using an exact context (i.e. the full sentence) for this purpose would result in extremely sparse and high-dimensional representations, a common simplifying assumption is to represent the context as a bag of words. Furthermore, since sentences in a corpus can be arbitrarily large, and most of the lexical unit's meaning is determined by its local context, a further assumption reduces the scope for bag-of-words extraction to a fixed window around the target word. The main drawback of pure count-based distributional representations is their inability to model the saliency of contexts. All contexts are treated equally, leading to over-representation of frequent non-discriminative contexts (like function words and auxiliary verbs) and under-representation of rare and highly specific contexts. As a solution, numerous association measures have been proposed as replacement for raw counts, e.g. *tf-idf*, *Dice coefficient*, *pointwise mutual* *information* and others; see e.g. (Navigli and Martelli, 2019) for a recent overview.

The framework of distributional semantics does not pose strict limitations on how the context is formulated. This enabled experiments with context definition: Lin (1998), highlighting the issues associated with linear context windows, proposes using syntactic triplets as context representation and demonstrates its efficiency for thesaurus extraction; Gabrilovich and Markovitch (2007) use Wikipedia article identifiers as context and suggest representing words based on their occurrence in corresponding Wikipedia pages, leading to improved word and text similarity performance.

The general advantage of sparse, explicit distributional semantic models is their simplicity and transparency: the dimensions of the resulting models are interpretable and the target-context association values can be traced back to the source corpora. However, such models suffer from sparsity as most targets and contexts never occur together, and from low efficiency, as most count-based association measures require the calculation of global, corpus-level statistics. These issues have been addressed by the next generation of distributional semantic models, the former by the models based on matrix factorization, and the latter by efficient neural network-based models that allow iterative refinement of the vector space. While our study is based on a more recent model, *word2vec*, our observations apply to the general distributional semantics setup and are thereby expected to hold for sparse distributional models as well.

2.2.3 Dense models and *word2vec*

Explicit matrices that map targets to contexts tend to be very large and very sparse: an association measure is provided for each target-context pair. However most targets and contexts rarely co-occur; explicit representations are susceptible to the noise in the data and do not take the semantically meaningful co-occurence regularities into account. A range of matrix factorization methods based on global co-occurence statistics have been proposed to address these issues, for example, Latent Semantic Analysis (LSA) (Landauer et al., 1998) uses singular value decomposition to represent the original association matrix as a product of lower-rank matrices, one of which contains a compact representation of the original vector space organised along the axes that reflect correlations between the columns and rows of the original matrix. LSA and related methods are computationally demanding and updating the resulting low-dimensional space based on new data is non-trivial. This has limited the amount of raw textual data that could be realistically ingested by these models, and it has been later shown that the ability to process large amounts of input data is key to model performance. The next generation of dense type-level distributional semantic models originated from the research in *language modeling* and provided much more efficient ways to pre-train high-quality word representations, enabling pre-training on huge, web-scale datasets and establishing the state of the art in static word representations that holds until today[3].

The goal of language modelling is to estimate the probability of a word given its context. Traditional non-neural language models represent words as discrete units and use n-grams – fixed-length sequences of preceding words – as context. The major drawbacks of n-gram language models are their rigidness when it comes to modelling word sequences not seen in the training data, and the sparse, high-dimensional nature of the learned representations, similar to the issues associated with explicit distributional word representations. While the former has been partially addressed by smoothing the probabilities and backing off to lower-order n-grams (Chen and Goodman, 1996), the latter posed a more fundamental challenge. The seminal work of Bengio et al. (2003) addressed both challenges by proposing a *neural* language model that associated each input word with a dense vector – a word embedding – and jointly learned the word representations and the sequence probability, leading to substantially improved performance and providing a principled way to address both out-of-domain and sparsity issues.

While Bengio et al. (2003) focused on language modeling per se, others explored the properties of the learned word representations. Collobert and Weston (2008) have introduced the all-neural end-to-end approach to natural language processing that has since then become mainstream methodology in the field. The key enabling factor in the success of end-to-end NLP is transfer learning: the representations learned for one (often, simpler and more data-rich) task can be re-used for solving another task. However, for Collobert and Weston (2008) transfer learning from scarce labeled data alone only yielded modest results, and most of the performance gain was achieved by transfer from pre-trained base word representations derived from unlabeled corpora based on the neural language modeling objective. The resulting pre-trained word embeddings have drastically improved the end-task performance, and a closer examination of the resulting dense representations has revealed semantic regularities in the learned space, e.g. grouping together the representations of country names, colours, videogame consoles, etc.

Despite the overall efficiency and simplicity compared to the NLP pipelines at the time, the word embedding model of Collobert and Weston (2008) was itself not very efficient and required weeks (Mikolov et al., 2013b) to train even on datasets that would be considered modestly-sized by modern standards. Neural pre-trained word embeddings have gained wide popularity with the release of an efficient implementation by Mikolov et al. (2013a) – *word2vec*. The proposed architecture employs a range of simplifying assumptions and tricks that allow pre-training of embedding models on huge unlabelled datasets within hours, and the authors demonstrate that given enough unlabelled data, the simplicity of the model is compensated by the amount of language material it is able to ingest. Word2Vec has become a standard pre-training approach for word embeddings, and has shown excellent results on a range of benchmarks,

[3] Vulić et al. (2020) have recently shown that lexical embeddings extracted from modern Transformer-based models *can* outperform type-level *fastText* representations (Bojanowski et al., 2017) in some evaluation settings, while still being much more expensive to train and to apply.

outperforming the neural and non-neural semantic models of that time by a significant margin. Since we chose word2vec as the algorithmic basis for our experiments, we briefly discuss the main building blocks of this model below.

Architecture. In terms of general architecture, word2vec follows the standard neural language model setup: each input word type is associated with a randomly initialized word embedding vector, which is trained via backpropagation with a language modeling objective. Unlike the models of Bengio et al. (2003) and Collobert and Weston (2008), the neural model of word2vec has no non-linear hidden layer, which reduces its ability to model complex interactions, but allows for much more efficient computation of word embeddings.

Models. In their original approach, Bengio et al. (2003) used sequence probability as an objective function. Collobert and Weston (2008) replaced it with a ranking objective, aiming to score correct word sequences above incorrect ones. Word2vec introduces two novel objectives which are easy to compute and allow taking not only previous, but also following words into account. Continuous bag-of-words (CBOW) aims to predict the target word type w_i given the context window of size m around its occurrence in a corpus, represented as an unordered bag of words. Continuous Skip-gram (SG), in turn, predicts each word in context based on the target word type w_i. The difference between the resulting architectures is depicted in Figure 2.3.

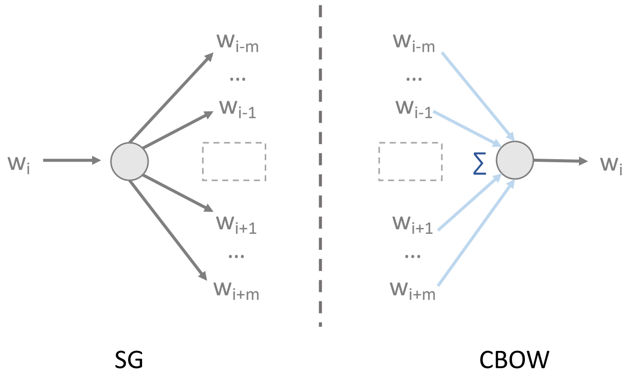


Figure 2.3: Word2vec architectures, based on Mikolov et al. (2013a).

Objectives, Sampling and Frequency Threshold. Word2vec employs a range of techniques to improve the training efficiency. One of the major sources of computational complexity in neural language models is the final projection layer, which in the default scenario contains activations for each and every word in the model's vocabulary. As an alternative to the hierarchical softmax objective (Morin and Bengio, 2005), Mikolov et al. (2013b) introduce the Negative Sampling (NS) where during training each word is only compared to a random k negative samples from the vocabulary. In addition, word2vec acquires large performance gains via subsampling of frequent words during training and employing *frequency thresholds* on target and context words to filter out noisy and extremely rare words. Following the results from (Mikolov et al., 2013b), Skip-Gram with Negative Sampling (SGNS) has become the most used word2vec variant.

The simplicity of the model, clever optimization strategies and efficient implementation have allowed word2vec to process much larger datasets within reasonable time frame, leading to state-of-the-art performance on word similarity and relatedness benchmarks. Most importantly for our work, the resulting word representations were shown to display a range of semantic and syntactic regularities, grouping similar words together and mirroring some semantic relationships between pairs of words within the vector space, as in the famous $w2v(king) - w2v(man) + w2v(woman) \approx w2v(queen)$ example.

Although some studies have demonstrated the systematic superiority of neural word embeddings (Baroni et al., 2014), it has been soon shown that their outstanding performance stems from their efficiency and flexible parametrization. In particular, Levy and Goldberg (2014a) demonstrate that the representation space produced by SGNS is an approximation of a factorized PMI matrix and that under comparable conditions a PMI-based distributional semantic models reach similar performance and semantic capacity; Levy et al. (2015) elaborate on this finding and conduct a large-scale comparison of popular distributional models, reporting that the differences between existing approaches can be largely attributed to the implicit hyperparameter choices and training data, showing no substantial performance differences between word2vec, PMI-based models and GloVe (Pennington et al., 2014). Quoting Levy and Goldberg (2014b), it turns out that "*neural embedding process is not discovering novel patterns, but rather is doing a remarkable job at preserving the patterns inherent in the word-context co-occurance matrix*".

Despite that, the efficiency and ease of use have made word2vec one of the default approaches for learning static word embeddings, and despite the recent developments in static and dynamic word representation research, plain word2vec models are still widely used in both mainstream NLP and interdisciplinary applications.

2.2.4 Limitations and Extensions

The classic word embedding approaches discussed above have several well-known limitations that have been addressed in subsequent work. We review these core limitations and and the proposed solutions below.

Unknown words. One of the core advantages of distributional semantic models is high coverage: since these models can be trained on large collections of unlabeled texts, one can easily obtain representations for every token encountered in the corpus without manual effort. Despite that, unknown token types are commonplace in word embeddings due to the domain shift, low frequency and morphological inflection. In addition, some token types are too rare to produce reliable representations for them, some are misspellings and noise, and some are preprocessing artifacts. To address the latter, frequency thresholding is applied to the vocabulary, reducing the size of the models – and, simultaneously, their coverage. While early models have used special tokens to represent rare words, a more principled approach to the unknown word issue is to learn representations for subword units instead of full token types. Prominent examples of type-level subword models include *Charagram* (Wieting et al., 2016) that represents each word as a sum of character ngrams, *fasttext* (Bojanowski et al., 2017) that extends the Skip-gram model with bag-of-ngrams word representation and *BPEemb* (Heinzerling and Strube, 2018) that employs byte-pair encoding (Sennrich et al., 2016) to learn meaningful sub-word units during training, and others. Subword embeddings can produce representations for rare and unknown words and achieve superior performance on rare word benchmarks, as well as in languages with complex inflectional morphology. Unlike their type-based predecessors, most subword models do not operate with linguistically meaningful units, which makes it challenging to inject linguistic information into the pre-trained vector spaces. To keep the experimental setup simple, we thereby fall back to type-based vector spaces, leaving the adaptation to subword-aware models for future work.

Lack of linguistic information. Traditional word embedding models like word2vec and GloVe have been trained on large collections of unprocessed raw text. Despite their ability to capture certain linguistic regularities (demonstrated by analogy task evaluation), it has been shown that introducing linguistic information into the models can greatly benefit their performance. At training time, Levy and Goldberg (2014c) replace linear context windows in word2vec SGNS with contexts based on syntactic dependencies, similar to (Lin, 1998). They show that the resulting embeddings better capture relational similarity between words. A parallel line of work has focused on word embedding targets: Ebert et al. (2016) introduce lemmatized LAMB embeddings and show that lemmatization is highly beneficial for word similarity and lexical modeling tasks even for morphologically poor English, while using *stems* doesn't lead to significant improvements. Finally, a line of research in retrofitting (Faruqui et al., 2015) aims to inject linguistic information into the already trained distributional model: Vulić et al. (2017b) use a small set of morphological rules to specialize the vector space bringing the forms of the same word closer together while setting the derivational antonyms further apart. Our study also falls into this category: we evaluate the impact of linguistic information injected at the training stage on the properties of the learned embeddings; we report results on window-based and dependency-based contexts, and – similar to (Ebert et al., 2016) and (Trask et al., 2015) – experiment with preprocessing the VSM targets at training time, generalizing and expanding on their initial findings.

Lexical ambiguity. Word embedding approaches like word2vec and GloVe produce a single representation for each token type. The obvious limitation associated with this is the inability of static word embeddings to reflect the polysemy of the natural language, conflating all the meanings associated with a particular type into a single representation. This issue has been addressed via multi-modal word embeddings (Athiwaratkun and Wilson, 2017) and Gaussian embeddings (Vilnis and McCallum, 2014). Partially disambiguating the source data via POS tagging has been employed in (Trask et al., 2015). Here, instead of constructing vectors for word forms, POS information is integrated into the vector space as well. This approach is similar to our `type.POS-w2` setup (Section 2.3) which, as we show, introduces additional sparsity and ignores morphological inflection. An alternative line of research aims to learn embeddings for lexical units by using an external word sense disambiguation (WSD) tool to preprocess the corpus and applying standard word embedding machinery to induce distributed representations for lexical units (Iacobacci et al., 2015; Flekova and Gurevych, 2016). Such approaches require an external WSD tool, which might introduce bias and be unfeasible for low-resource languages and domains. Moreover, to query such VSMs it is necessary to either apply WSD to the input, or to align the inputs with the senses in some other way, which is not always feasible. While not addressing the ambiguity issue directly, in this work we present several ambiguity-aware resource-based evaluation scenarios for static word embeddings and demonstrate that even partial disambiguation on the grammatical level leads to better alignment between the learned word representations and the lexical evaluation benchmarks.

Context-independence. A related challenge associated with static word representations is their inability to adapt the learned representations to the context. Taking context into account implicitly addresses the lexical ambiguity, and modern contextualized word representations like BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) show remarkable sense disambiguation capabilities without task-specific training. While we discuss contextualized encoders in much more detail in Chapter 3, it is important to reiterate that despite the disadvantages, context-independence and low computational cost are desirable properties in many applied and experimental setups, and static word embeddings like word2vec are widely used in knowledge graph representations (Sorokin and Gurevych, 2018), bias research (Sweeney and Najafian, 2019) and interdisciplinary applications (Hartman et al., 2020; Wang et al., 2020).

Opaqueness. Finally, the ability to ingest large datasets and efficiently represent complex distributional relationships comes at the price of transparency. While explicit word representations based on raw co-occurence data are sparse and inefficient to train and to apply, they allow direct inspection of the resulting models, which has been used to interpret the representations encoded by these models (Levy and Goldberg, 2014c; Lin, 1998). Such direct inspection is not possible in case of dense neural models, and alternative approaches have been proposed that allow some insight, e.g. compressing the resulting vector spaces into an easy-to-visualize dimensionality via matrix factorization techniques like t-SNE (van der Maaten and Hinton, 2008), examining the clusters emerging from the vector space models and analyzing the contexts activated by certain targets (Levy and Goldberg, 2014c). The lack of transparency remains an open issue for dense neural network-based models in general and is addressed by a rapidly growing line of research in probing, which we review in great detail in Chapter 3.

2.2.5 Evaluation of word embeddings

As demonstrated above, there exists a multitude of approaches to train distributional word representations, and many of them allow rapidly producing new vector space models from raw text at low cost. The ever-growing variety of pre-trained vector spaces calls for a solid evaluation and analysis methodology that would allow comparison of these models in terms of performance, and analysis of their qualitative properties.

Before we proceed, it is important to reflect on what in fact is the subject of word embedding evaluation. Benchmark performance has been historically used to demonstrate the superiority of one approach over the other, however, it has lately been shown that the measured differences in performance are largely due to a range of external factors that have little to do with the embedding model itself: the training data, the preprocessing and the hyperparameter choices. As a well-known example, the word2vec models have been claimed superior to PMI factorization (Baroni et al., 2014) both later outperformed by the GloVe embeddings (Pennington et al., 2014). However, Levy et al. (2015) show that under fair conditions, incl. optimal hyperparameter values and shared training data, the three approaches perform on-par, with PMI factorization having an upper hand in some cases. It is therefore important to remember that the subject of evaluation – independent of the chosen method – is always a *particular instance* of a model, produced from a certain collection of texts under a certain set of hyperparameters, and any claims about the impact of modeling and hyperparameter choices should be backed up by an experimental setup which isolates the independent variable and keeps the rest fixed. An alternative, practical line of action is to determine a model instance well suited for a downstream task (e.g. *"pre-trained dependency-based 300-dim word2vec from Levy, 2014"*), however, in that case any general claims about the word embedding method itself should be avoided. Our work takes the former stance and focuses on the evaluation of the impact of specific preprocessing choices on the intrinsic performance while keeping the rest of the setup fixed. With that in mind, we now turn to the main principled approaches to static word embedding evaluation.

Extrinsic evaluation. As mentioned above, the most practical, end-to-end approach for evaluating word embedding models is to measure their impact on downstream task performance. Dense neural models pre-trained on large text collections have indeed been shown to drastically improve end-task performance for most NLP tasks; the impact of the pre-trained model itself, however, can be obscured by the downstream task-specific components, the properties of the end task, the underlying data and domain mismatch, and the particularities of the task-specific evaluation.

Similarity benchmarking. An alternative, task-neutral way to evaluate and compare word embedding models is similarity benchmarking. A similarity benchmark is a set of word pairs labeled with association scores by human annotators. These scores are compared to the similarity between the respective words vectors in the pre-trained vector space model. High correlation between the vector similarities and human-labeled association scores signals that the target relation is encoded is the models' geometry. Pre-trained vector space models can then be compared in terms of their correlation scores. While early similarity benchmarks have been criticised for the subjectivity of human label assignments, this approach has remained popular due to its simplicity, speed, transparency and ease of interpretation. We re-visit similarity benchmarks in Section 2.4 for a more in-depth discussion.

Analogy tasks. Traditional word pair-based benchmarks measure attributional similarity between their items, e.g. a *car* would be similar to a *bus*, but less similar to a *flower*. It has been shown that pre-trained distributional representations also encode relational similarity between pairs of words that stand in a certain semantic relation (Turney, 2006). An example of relational similarity is the seminal *"a king to a man is the same as a queen to a woman"*. The notion of relational similarity has been captured in a range of word analogy datasets, incl. the Google Analogy test set (Mikolov et al., 2013a) and BATS (Gladkova et al., 2016). An analogy task consists of a word pair and a query word, as well as a set of candidate fillers; the objective is to select an appropriate candidate filler based on the example word pair and the query. Despite their intuitive appeal, word analogy tests have been criticized for a number of theoretical and methodological drawbacks (Rogers et al., 2017). While we do not investigate the analogy task performance in this work, we note that our resource-based evaluation scenario can be cast as a special case of an analogy task.

Resource-based evaluation. As a further step towards similarity benchmark specificity, one can employ existing lexical-semantic resources and investigate what semantic relations and groupings from these resources are encoded by the embedding model under investigation. Once constructed, expert-curated LSRs can be used to extract a wide range of principled, well-defined tests for static word embeddings; however, resource-based evaluation naturally faces the coverage issues on the resource side. The word class suggestion task that we introduce later in this chapter falls into this category, and we discuss our resource-based evaluation in more detail in Section 2.5.

Similarity benchmarks as probes. Despite their use as task-neutral leaderboards for word embedding models, intrinsic evaluation methods for static word embeddings only cover certain aspects of language and turn out to correlate poorly with end-task performance (Faruqui et al., 2016). They are, however, useful for another reason: many of the intrinsic evaluation tasks for static word embeddings are in fact probing tasks: given a particular linguistic phenomenon (e.g. hyponymy); a frozen encoder (e.g. pre-trained word2vec) and a simple transformation (e.g. cosine similarity), a probing task measures the performance of the input encoder to discover what aspects of language can be easily extracted from the model. The core goal of intrinsic evaluation and probing is not to determine the best model for end applications, but to study the behavior of the model with respect to a set of isolated linguistic tasks, and to measure the effects of interventions (e.g. hyperparameter values) on this behavior. Our thesis embraces this stance: in what follows below, we do not aim to determine the *"best"* model, and instead focus on the effects of different model configurations measured in terms of probing task performance.

2.3 General setup

Having covered the background necessary to position our work, we turn to our experiments. The core observation of our study is that there exists a conceptual discrepancy between the vocabulary items of pre-trained vector space representations (token type) and the items used as targets in during word embedding evaluation (lexical unit). We hypothesize that this conceptual gap can lead to decreased performance and less semantically plausible vector space representations. To investigate this hypothesis, we are going to compare the performance of different vector space model configurations in two settings: similarity benchmarking will allow us to get a first estimate of the target effects and put our work into context of related research; the novel resource-based word class suggestion scenario will enable detailed insights into the lexical-semantic properties of the pre-trained representation spaces.

Distributional word representations aim to encode word targets based on the context words they co-occur with. Since our experimental setup will require training numerous configurations of the word embedding model from scratch, we chose the efficient SGNS from word2vec as our target model; however, since the hypothesized conceptual gap pertains to the evaluation setup and not to a particular model, we expect the results to generalize to any representation that operates with surface word types as targets.

In the original SGNS, targets and contexts belong to the same type-based vocabulary, but this is not required by the model. In this study we experiment with the following targets: `type` (*"going"*), `lemma` (*"go"*), `type.POS`[4] (*"going.V"*) and `lemma.POS` (*"go.V"*). Word embeddings demonstrate qualitative differences depending on context definition (Levy and Goldberg, 2014c),

[4] We use coarse POS classes by selecting the first character of the predicted Penn POS tags.

and we additionally report the results on 2-word window-based (`w2`) and dependency-based (`dep-W`) contexts. To keep the evaluation setup simple, we only experiment with type-based contexts.

To isolate the target of our study, for each setting we train a 300-dimensional VSMs on a 2018 English Wikipedia dump, preprocessed using the Stanford Core NLP pipeline (Manning et al., 2014) with Universal Dependencies 1.0 as syntactic formalism. The text is extracted using the *wikiextractor* module[5] with minor additional cleanup routines. Training the VSM is performed with the skip-gram based *word2vecf* (sic) implementation by Levy and Goldberg (2014c) with default algorithm parameters. The only variables that distinguish the resulting VSMs are the choice of target and the context type; the dimensionality, hyperparameters and underlying text data remain the same across experimental settings.

2.4 Similarity Benchmarking

2.4.1 Benchmark Design and Requirements

A similarity benchmark usually consists of a set of word pairs associated with human-labeled scores. These scores are compared to the cosine similarity outputs of a vector space model in terms of correlation score, higher correlation meaning higher degree of correspondence between human notion of similarity and the automatic vector encodings induced from pre-training data.

Despite the apparent simplicity, similarity benchmarks widely differ in the methodology applied during the data collection and annotation. The design choices during the similarity benchmark construction, in turn, affect its utility and validity as a measurement tool for word embedding performance. Below we describe the core design choices in similarity benchmark construction and the associated requirements (R) in the context of our study.

Similarity and Relatedness. The first core aspect of benchmark construction is the annotation target: what are the scores assigned by human annotators supposed to reflect? An important distinction made in literature is between functional similarity and topical relatedness. Functional similarity implies the structural, lexical similarity between compared items and covers phenomena such as synonymy, hypo- and hypernymy. Topical relatedness, on the other hand, does not imply any lexical-semantic relationship between the items, and amounts to the items standing in a broad co-occurrence relationship. As an illustraction, the words *"money"* and *"cash"* are functionally similar, while the words *"money"* and *"bank"* are not similar, but clearly related. Both similarity and relatedness have applications in natural language processing; however, the failure to strictly differentiate between the two has led to major criticism of the early similarity benchmarks. Since we are primarily

[5] `https://github.com/attardi/wikiextractor`

concerned with the reflection of lexical-semantic properties in the vector space models, we require our benchmark to strictly reflect functional similarity (R1).

Composition. The choice of words and the word pairing strategy can have major impact on the resulting dataset. It has been shown that the differences in *parts of speech* affect both the annotator agreement and the benchmarking scores – however, most early similarity benchmarks have focused on nouns. To reflect a broader spectrum of lexical-semantic phenomena, we thereby require a benchmark that contains a balanced selection of words from different parts of speech (R2). Moreover, since our study utilizes the part-of-speech information, the part of speech has to be explicitly annotated in the benchmark (R3). Another important dimension in similarity benchmark construction is the inclusion of *multi-word expressions* and *named entities*. Since here we focus on general lexical-semantic relationships between common words, however, we limit our focus to the benchmarks that use single common words as targets. Additional design choices like word rarity and lexical-semantic selection criteria can qualitatively affect the benchmarks as well; however, we do not impose any filtering restriction based on these.

Annotators and subjectivity. Even with the similarity/relatedness distinction ruled out, the human annotation task remains subjective, and several principled strategies have been proposed to mitigate the subjectivity and produce reliable scores. While earlier benchmarks have employed expert linguists and shown high inter-annotator agreement, this has posed limitations on dataset size. Alternative, crowdsourcing-based solutions have allowed the construction of larger benchmarking sets at the cost of lower agreement. The promising middle ground is to employ crowdsourcing coupled with demographic data and an expert in the loop. For example, to construct the SimVerb dataset Gerz et al. (2016) used the Prolific Academic platform to select annotators from specific demographics likely to produce consistent similarity rankings, e.g. native English speakers of age 18-50 born and residing in the US, UK or Ireland, leading to very good agreement scores in comparison to existing benchmarks. Naturally, in this work we are interested in benchmarks with high inter-annotator agreement (R4).

Size and complexity. Last but by far not least, the size of the dataset and the complexity of the covered phenomena are important features that determine the utility of a similarity benchmark. A modestly sized similarity benchmark might not represent the target phenomena well. A biased benchmark, in turn, might lead to inflated performance scores. Coupled with low inter-annotator agreement, this allows the vector space models to seemingly surpass human performance, making the future utility of the said benchmarks unclear. Thereby in our study we are interested in a reasonably large benchmark dataset covering a wide range of semantic phenomena (R5).

2.4.2 Dataset selection

benchmark	task (R1)	POS (R2/3)	workforce	IAA ρ (R4)	#pairs (R5)
RG-65	**sim**	N	expert	-	65
WS-353	rel	N	expert	0.61	353
MEN	rel	**NVA**	crowd	0.68	**3000**
RW	**sim**	NVA*	crowd	-	**2000**
SimLex	**sim**	**NVA**	crowd	0.67	**999**
SemEval-2017	**sim**	N*	expert	**0.9**	500
SimVerb	**sim**	**V**	crowd*	**0.85**	**3500**

Table 2.1: Benchmark overview, satisfied criteria and datasets selected for the study in bold. *: RW does not provide explicit POS tags. SimVerb has been annotated by crowdworkers pre-filtered with respect to demographics. SemEval-2017 containts nouns, multi-word expressions and named entities.

Having discussed the critical dimensions in similarity benchmark design, we now review several popular similarity benchmarks and position them in this space. Table 2.1 provides an overview of the benchmark sets considered here.

RG-65 Arguably the first similarity benchmark, RG-65 has been introduced by (Rubenstein and Goodenough, 1965) to empirically evaluate Harris' distributional hypothesis (Harris, 1954). While Hill et al. (2015) note that RG-65 indeed seems to encode similarity rather than relatedness, the benchmark only covers 65 word pairs, which limits its applicability for vector space model evaluation.

WS-353 A larger benchmark suggested by Finkelstein et al. (2001) has been a long-standing standard in the evaluation of distributed word representations. WS-353 contains 353 word pairs annotated by 16 annotators. It only covers nouns, and has been criticised for the lack of clear distinction between similarity and relatedness, tending to represent the latter (Faruqui et al., 2016), as well as for low inter-annotator agreement (Hill et al., 2015). An intermediate solution to this issue has been proposed by Agirre et al. (2009) who split the dataset into separate similarity and relatedness sections; however, the similarity subset is naturally smaller, and keeps the scores produced by WS-353 annotators, potentially compromising the evaluation in favor of relatedness.

MEN Targeting the evaluation of multi-modal word representations, the MEN dataset (Bruni et al., 2012) is an order of magnitude larger than WS-353, covering 3000 pairs of randomly selected words. The scores have been assigned via Amazon Mechanical Turk, and the vision-related metadata allows to gain insights into multi-modal properties of word representations. As with WS-353, the dataset has been criticised for the lack of distinction between similarity and relatedness, as well as for the bias towards concrete nouns stemming from the vision-driven data selection protocol (Hill et al., 2015).

RareWords (RW) Aiming to provide insights into the performance of pre-trained vector models for low-frequency phenomena, the RareWords dataset (Luong et al., 2013) covers over 2000 word pairs annotated by 10 raters via Amazon Mechanical Turk. The benchmark is composed of infrequent words and word forms, mostly nouns, and does not provide explicit part-of-speech annotations.

SimLex-999 To address the shortcomings of the existing word similarity benchmarks, SimLex-999 (Hill et al., 2015) covers a balanced selection of 999 words classified by part of speech and accompanied with lexical-semantic relations extracted from WordNet. Unlike WS-353 and MEN, the dataset explicitly focuses on similarity (as opposed to relatedness), only pairs together words from the same part of speech, and provides a balanced selection of concrete and abstract words. The dataset was annotated via Amazon Mechanical Turk, leading to agreement scores comparable to the ones of WS-353 and MEN (Spearman $\rho = 0.67$, 0.61 and 0.68 respectively). The focus on similarity, the wide POS coverage and the explicit POS annotation make SimLex-999 a valid target for our study. Subsequent work (Camacho-Collados et al., 2017) has criticised this dataset for moderate agreement and lexical-semantic biases in item selection, which leads to over-representation of antonyms and makes it possible to inflate the benchmark performance by explicit antonymy modeling. We acknowledge these drawbacks and point out that the inter-annotator agreement of SimLex-999 is still within the common range for similarity benchmarks. We indeed encounter traces of the reported antonymy bias while comparing to related work (Vulić et al., 2017b); however, since our intervention targets purely inflectional phenomena, the effects of this bias are orthogonal to our findings.

SemEval 2017 Targeting multi-lingual word embedding evaluation, the dataset of SemEval Shared Task 2 (Camacho-Collados et al., 2017) covers 500 English word pairs diversified via BabelDomains (Camacho-Collados and Navigli, 2017) to increase the domain coverage. The data was annotated in a generative manner, where the annotators were asked to suggest a word given an input and a target similarity rating, followed by independent score annotation. The dataset includes a large proportion of named entities and multi-word expressions, which falls outside of the scope of our work, since we focus on common single-word entries.

SimVerb-3500 Noting the importance of verb semantics and the lack of dedicated evaluation resources, SimVerb-3500 (Gerz et al., 2016) provides a large-scale dataset of 3500 verb pairs sampled by VerbNet class and annotated by demographically filtered crowdworker pool. Arguably the most advanced similarity benchmark out of the ones discussed here, it explicitly targets similarity, has high coverage coupled with good inter-annotator agreement levels ($\rho = 0.84 - 0.86$), focuses on single-word entries and aligns well with the overlapping SimLex annotators. These qualities, along with our later focus on verb-based predicate semantics, make SimVerb an ideal target for our study.

2.4.3 Experiment

Based on the outlined criteria, we have selected SimLex-999 and SimVerb-3500 as two target benchmarks. Both benchmarks explicitly focus on similarity as opposed to relatedness, provide wide coverage stratified by part of speech and explicitly mark POS for the items, which is required by our POS-enriched VSMs. SimLex-999 contains nouns (60%), verbs (30%) and adjectives (10%); SimVerb-3500 is verbs-only. We evaluate eight vector space models: four targets (`type`, `type.POS`, `lemma` and `lemma.POS`) in two context configurations (`w2` and `dep`), as defined in Section 2.3.

One important observation pertains to the vocabulary coverage of the resulting VSMs. The vocabulary coverage of a particular VSM instance given the same corpus and frequency threshold (discussed in Section 2.2.3) depends on how the targets are defined. Table 2.2 provides coverage statistics for our similarity benchmarks. A notable effect of POS disambiguation is *vocabulary fragmentation*, when the same word type is split into several entries based on the POS, e.g. *acts (110) → acts.V (80) + acts.N (30)*. As a result, some word types do not surpass the frequency threshold and are not included into the final VSM. This effect is partially compensated by *lemma-based normalization*, which merges the word forms and increases the target count again, allowing more targets to pass the threshold.

	SimLex				SimVerb
VSM target	N	V	A	all	V
type	100.00	99.41	100.0	99.90	99.27
+ POS	99.20	99.41	100.0	99.32	90.08
lemma	100.0	100.0	100.0	100.0	99.76
+ POS	100.0	100.0	100.0	100.0	99.40

Table 2.2: Similarity benchmark coverage (%)

2.4.4 Results

Table 2.3 summarizes the performance of the VSMs in question on the selected similarity benchmarks.

Several observations can be made. Lemmatized targets generally perform better, with the boost being more pronounced on SimVerb. English verbs have richer morphology than other parts of speech and benefit more from lemmatization. Adding POS information benefits the SimVerb and SimLex verb performance, which can be attributed to the coarse disambiguation of verb-noun and verb-adjective homonyms. The `type.POS` targets show a considerable performance drop on SimVerb and SimLex verbs due to vocabulary fragmentation,

Context: `w2`

	SimLex				SimVerb
target	N	V	A	all	V
type	.334	**.336**	**.518**	.348	.307
+ POS	.342	.323	.513	.350	.279
lemma	**.362**	.333	.497	**.351**	.400
+ POS	.354	**.336**	.504	.345	**.406**
* type	-	-	-	.339	.277
* type MFit-A	-	-	-	.385	-
* type MFit-AR	-	-	-	.439	.381
Context: `dep-W`					
type	.366	.365	.489	.362	.314
+ POS	.364	.351	.482	.359	.287
lemma	**.391**	.380	**.522**	**.379**	.401
+ POS	.384	**.388**	.480	.366	**.431**
* type	-	-	-	.376	.313
* type MFit-AR	-	-	-	.434	.418

Table 2.3: Benchmark performance, Spearman's ρ. Results marked with * are the SGNS results from Vulić et al. (2017b); **bold** – best result among our models for a given context definition; underline – best result among our models overall.

compensated by lemmatization in `lemma.POS`. Using `dep` contexts proves beneficial for both datasets since modeling the context via syntactic dependencies results in more similarity-driven (as opposed to relatedness-driven) word embeddings (Levy and Goldberg, 2014c).

We provide the Morph-Fitting scores (Vulić et al., 2017b) for reference; a direct comparison is not possible due to the differences in the training data and the information available to the models. Vulić et al. (2017b) use word type-based VSMs specialized via Morph-Fitting (`MFit`), which can be seen as an alternative to lemmatization. Morph-Fitting consists of two stages: the Attract (`A`) stage brings word forms of the same word closer in the VSM, while the Repel (`R`) stage sets the derivational antonyms further apart. Lemma grouping is similar to the Attract stage. However, comparing the `MFit-A` and `-AR` results reveals that a major part of the Morph-Fitting performance gain on SimLex comes from the derivational Repel stage[6], which is out of the scope of our approach and plays into the aforementioned dataset bias of SimLex towards over-representation of antonyms.

While some properties of lemmatized and POS disambiguated embeddings are visible on the similarity benchmarks, the results are inconclusive, and we proceed to a more detailed evaluation scenario.

[6] See also (Vulić et al., 2017b), Table 5.

2.5 Word Class Suggestion

2.5.1 Resource-based Evaluation of Word Embeddings

Similarity benchmarks serve as a standard evaluation tool for word embeddings, but provide little insight into the nature of relationships encoded by the word representations. A more fine-grained context-free evaluation strategy is to assess how well the relationships in a certain *lexical resource* are represented by the given VSM. Two general approaches to achieve this are rank-based and clustering-based evaluation.

Rank-based evaluation treats the lexical resource as a graph with entries as nodes and lexical relations as edges, and estimates how well the similarities between the VSM targets represent the distances in this graph via mean reciprocal rank (MRR); for example, the related work of Ebert et al. (2016) uses WordNet (Miller, 1995) for this purpose. Rank-based evaluation requires the target lexical resource to have a dense linked structure which might not be present.

Clustering-based evaluation groups the entries of a VSM into clusters and compares these clusters to meaningful groupings extracted from a lexical resource. Vulić et al. (2017a) utilize the VSM to produce target clusters which are compared to the groupings from the lexical resource via collocation and purity. This approach only requires lexical entries to be grouped into classes and does not make assumptions about the density of the resource structure. However, clustering-based evaluation doesn't account for word ambiguity: a word can only be assigned to a single cluster. Moreover, using an external parametrized clustering algorithm introduces an additional level of complexity which might obscure the VSM performance details.

Odd Man Out is a promising evaluation strategy on the intersection of analogy and resource-based evaluation that has been suggested in work concurrent to our study (Stanovsky and Hopkins, 2018). The authors propose to generate odd-man-out puzzles consisting of a set of words with one outlier, e.g. *orange*, *apple*, *lemon*, *potato*. They introduce a crowdsourcing-based setup for puzzle generation and use the resulting datasets to evaluate a range of resource-based solvers as well several word embedding methods – including a polysemous representation induced from ELMo embeddings via clustering. While the puzzles suggested by Stanovsky and Hopkins (2018) are constructed by the crowd, adapting their evaluation setup for pure resource-based scenario could be an interesting expansion of our method described below.

We propose a suggestion-based evaluation approach to word embedding evaluation: we use the source VSM directly to generate *word class suggestions* (WCS) for a given input term. Many lexical resources group words into intersecting word classes, providing a compact way to describe word properties on the class level. For example, in VerbNet (Schuler, 2005) verbs can belong to one or more Levin classes (Levin, 1993) based on their syntactic behavior, FrameNet (Baker et al., 1998) groups its entries by the semantic frames they can evoke, and WordNet (Miller, 1995) provides coarse-grained supersense groupings. Suggestion-based evaluation can be seen as a flexible alternative to clustering-based evaluation, which intrinsically takes ambiguity into account and does not require an additional clustering layer. The following section describes the suggestion-based evaluation of static word embeddings in more detail[7].

2.5.2 Methodology

Task formulation

Abstracting away from the resource specifics, a *lexicon* L defines a mapping from a set of lexicon *members* $m_1, m_2, ...m_i \in M$ to a set of *word classes* $c_1, c_2...c_j \in C$. We denote the set of classes available for a member m as $L(m)$ and the set of members for a given class c as $L'(c)$. Given a *query* q, our task is to provide a set of word class suggestions $S_L(q) = \{c_a, c_b...\} \in C$. Note that we aim to predict all potential classes for a member given its vector representation on vocabulary level, independent of context.

Suggestion strategies

Given an input target w, the vector space model V, in turn, provides its vector representation $V(w)$. We use a measure of similarity between the vector representations of targets $sim(V(w_i), V(w_j))$ to rank the word classes. In this work we use cosine similarity.

A lexical resource might already contain a substantial number of members, and a natural strategy for word class suggestion is to find the *prototype* member m_{proto} closest to the query q in the VSM, and use its classes as suggestions. This scenario mimics human interaction with the lexicon. If $q \in M$, this is equivalent to a lexicon lookup. More formally,

$$m_{proto} = argmax_{m \in M} sim(V(q), V(m))$$

$$score_{proto}(q, c, L) = \begin{cases} 1, & \text{if } c \in L(m_{proto}) \\ 0, & \text{otherwise} \end{cases}$$

The prototype strategy per se is sensitive to the coverage gaps in the lexicon and inconsistencies in the VSM. We generalize it by ranking each word class $c \in C$ using the similarity between the query q and its closest member in c:

$$score_{top}(q, c, L) = max_{m \in L'(c)} sim(V(q), V(m))$$

This is equivalent to performing the prototype search on each word class, and scoring each class by the closest prototype among its members, given the query.

[7] Our implementation of WCS is available at https://github.com/UKPLab/coling2018-wcs

We use this generalized strategy in our further experiments. The output of the word class suggestion model $S_L(q, V)$ is a set of classes ranked using the input VSM, as illustrated by Table 2.4.

query	**top classes / prototypes**
dog→	animal (*cat*), food (*rabbit*) ...
crane→	artifact (*derrick*), animal (*skimmer*) ...
idea→	cognition (*concept*), artifact (*notion*) ...
bug→	animal (*worm*), state (*flaw*) ...

Table 2.4: WCS output for WordNet supersenses

Evaluation procedure

For each member m in the lexicon in turn, we remove it from the lexicon, resulting in a reduced lexicon L_{-m}. We aim to reconstruct its classes using the suggestion algorithm and the remaining mappings.

The performance is measured via precision (P) and recall (R) at rank k with the list of original classes for a given lexical unit serving as ground truth. Formally, given m, we compute the ranking $S_{L_{-m}}(m, V)$. Let $S_{@k}$ be the set of classes suggested up to the rank k, and T be the true set of classes for a given member in the original lexicon. Then

$$P_{@k} = \frac{|S_{@k} \cap T|}{|S_{@k}|} \qquad R_{@k} = \frac{|S_{@k} \cap T|}{|T|}$$

To get a single score, we average individual members' $P_{@k}$ and $R_{@k}$ for each value of k, resulting in scores $\overline{P}_{@k}$ and $\overline{R}_{@k}$. F-measure might be then calculated using the standard formula

$$F_{@k} = \frac{2\overline{P}_{@k}\overline{R}_{@k}}{\overline{P}_{@k} + \overline{R}_{@k}}$$

Upper bound Since the number of gold classes is not known in advance, the evaluation is always performed on k ranks, which leads to a resource-specific upper bound on P. For example, if a member only has one class, the ranked list of 10 suggestions will inevitably show lower precision. When the member set is not fully covered by the VSM target vocabulary, an additional upper bound on R applies.

2.5.3 Lexical resources

We use two lexical resources for suggestion-based evaluation: VerbNet 3.3 (Schuler, 2005) and WordNet 3.1 (Miller, 1995). VerbNet groups verbs into classes[8] so that verbs in the same class share syntactic behavior, predicate

[8] For simplicity, in this study we ignore subclass divisions.

semantics, semantic roles and restrictions imposed on these roles. For example, the verb *"buy"* belongs to the class *get-13.5.1*. The class specifies a set of available roles, e.g. an animate `Agent` (buyer), an `Asset` (price paid) and a `Theme` (thing bought), and lists available syntactic constructions, e.g. the `Asset V Theme` construction (*"$50 won't buy a dress"*). A verb might appear in several classes, indicating different verb senses. For example, the verb *"hit"* allows several readings: as *"hurt"* (*"John hit his leg"*), as *"throw"* (*"John hit Mary the ball"*) and as *"bump"* (*"The cart hit against the wall"*). VerbNet has been successfully used to support semantic role labeling (Giuglea and Moschitti, 2006), information extraction (Mausam et al., 2012) and semantic parsing (Shi and Mihalcea, 2005).

WordNet, besides providing a dense network of lexical relations, groups its entries into coarse-grained supersense classes, e.g. `noun.animal` (*"aardvark"*, *"koala"*), `noun.location` (*"park"*, *"senegal"*), `noun.time` (*"forties"*, *"nanosecond"*). WordNet supersense tags have been applied to a range of downstream tasks, e.g. metaphor identification and sentiment polarity classification (Flekova and Gurevych, 2016). WordNet differs from VerbNet in terms of granularity, member-class distribution and part of speech coverage, and allows us to estimate VSM performance on nominal as well as verbal supersenses, which we evaluate separately. Table 2.5 provides the statistics for the resources. We henceforth denote VerbNet as `VN`, WordNet nominal supersense lexicon as `WN-N` and WordNet verbal supersense lexicon as `WN-V`.

	classes	members	ambig	%ambig
VN	329	4 569	1 366	30%
WN-V	15	8 702	3 326	38%
WN-N	26	57 616	9 907	17%

Table 2.5: Lexicon statistics, single-word members

2.5.4 Experiment

We use the suggestion-based evaluation to examine the effect of lemmatization and POS-disambiguation using the same eight VSM configurations as before. The coverage analysis of the VSMs and lexica is presented in Table 2.6. For brevity, we only report coverage on `w2` contexts. We have observed slight coverage differences for `dep` contexts, and attribute this to the context vocabulary fragmentation caused by dependency typing of the contexts, similar to the POS fragmentation effect described earlier.

Coverage analysis on lexica confirms our previous observations: lemmatization allows more targets to exceed the SGNS frequency threshold, which results in consistently better coverage. POS-disambiguation, in turn, fragments the vocabulary and consistently reduces the coverage with the effect being less pronounced for lemmatized targets. `WN-N` shows low coverage containing many low-frequency items. Due to the significant discrepancies in VSM coverage, we

target	VN	WN-V	WN-N
type	81	66	47
+POS	54	39	43
lemma	88	76	53
+POS	79	63	50
shared	54	39	41

Table 2.6: Lexicon member coverage (%)

conduct our experiments on *shared vocabulary*, only including members found in all VSMs to analyze the qualitative differences between VSMs.

2.5.5 Results

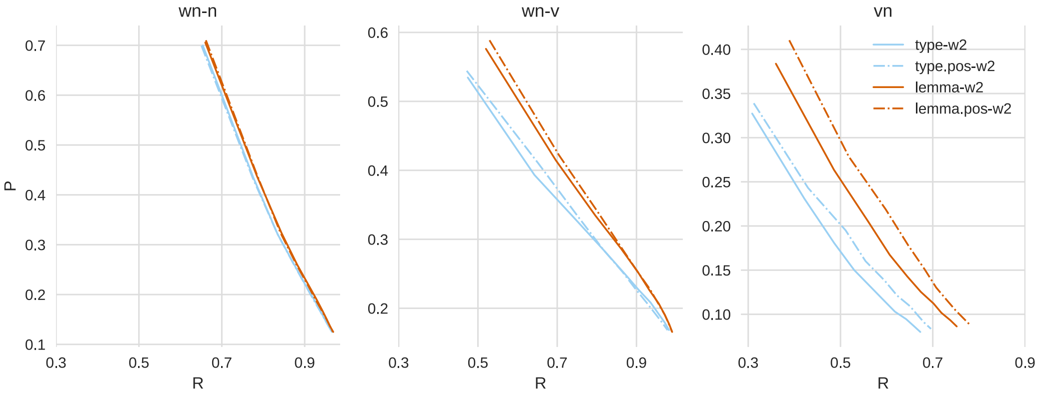


Figure 2.4: WCS PR-curve, shared vocabulary, `w2` contexts.

We treat the cutoff rank k as a parameter that specifies the Precision-Recall trade-off. As Figure 2.4 demonstrates, lemmatized targets consistently outperform their word form-based counterparts on the WCS task. The magnitude of improvements varies between resources: verb-based `WN-V` and `VN` benefit more from lemmatization, and `VN` gains most from POS-disambiguation. This aligns with the similarity benchmarking results where the verb-based SimVerb benefits more from the addition of lemma and POS information.

Table 2.7 provides exact scores for reference. Note that the shared vocabulary setup puts the `type` and `type.POS` VSMs at an advantage since it eliminates the effect of low coverage. Still, `lemma`-based targets significantly[9] ($p \leq .005$) outperform `type`-based targets in terms of F-measure in all cases. For window-based `w2` contexts POS disambiguation yields significantly better F scores on lemmatized targets for `VN` ($p \leq .005$) with borderline significance for `WN-N` and `WN-V` ($p \approx .05$). When dependency-based `dep` contexts are used, the effect

[9] Wilcoxon signed-rank test over individual lexicon members' F scores

	WN-N			WN-V			VN		
	P	R	F	P	R	F	P	R	F
Context: `w2`									
type	.700	.654	.676	.535	.474	.503	.327	.309	.318
+POS	.699	.651	.674	.544	.472	.505	.339	.312	.325
lemma	.706	.660	.682	.576	.520	.547	.384	.360	.371
+POS	**.710**	**.662**	**.685**	**.589**	**.529**	**.557**	**.410**	**.389**	**.399**
Context: `dep`									
type	.712	.661	.686	.545	.457	.497	.324	.296	.310
+POS	.715	.659	.686	.560	.464	.508	.349	.320	.334
lemma	**.725**	**.668**	**.696**	.591	.512	.548	.408	.371	.388
+POS	.722	.666	.693	**.609**	**.527**	**.565**	**.412**	**.381**	**.396**

Table 2.7: WCS performance, shared vocabulary, $k = 1$. Best results across VSMs in bold.

of POS disambiguation is only statistically significant on `type` targets for `VN` ($p \leq .005$) and on `lemma`-based targets for `WN-V` ($p \leq .005$). We attribute this to the fact that dependency relations used by `dep` contexts are highly POS-specific, reducing the effect of explicit disambiguation. Lemma-based targets without POS disambiguation perform best on `WN-N` when dependency-based contexts are used; however, the difference from lemmatized *and* disambiguated targets is not statistically significant ($p > .1$).

	SimLex			SimVerb	WN-N	WN-V	VN
	N	V	A	V	N	V	V
base	.80	.26	1.0	.24	.86	.21	.22
avg #POS	1.08	1.01	1.39	1.50	1.15	1.37	1.42
single POS	.93	.99	.62	.51	.85	.65	.59

Table 2.8: Base form ratio and available POS averaged over members; % members with single POS.

Our results are in line with the previous observations (Gerz et al., 2016) in that the verb similarity is hard to capture with standard VSMs. To investigate why verbs benefit most from lemmatization and POS disambiguation, we analyze some relevant statistics based on the Wikipedia data that we have used to train our models. Table 2.8 shows the ratio of base form to total occurrences in our corpus[10]. As we can see, the base form (lemma) is by far not the dominating form for verbs. Practically this means that for our resources verbal `type` targets have direct access to only 20-25% of the corpus occurrence data on average. Individual verbs and nouns also differ in terms of the frequency distribution of their word forms, which introduces an additional bias into the evaluation. This effect is countered by lemmatization.

[10] We use all lemmas that appear more than 100 times in the corpus; to smooth the effect of tagging errors we only count POS that appear with the target lemma in more than 10% of total lemma occurrences. All statistics averaged over individual lemmas.

The second important difference between the noun- and verb-based lexica is the number of POS available for the lexicon members' lemmas. As Table 2.8 further demonstrates, nouns are less ambiguous in terms of POS: for example, `VN` member lemmas appear with 1.42 distinct POS categories on average, compared to 1.15 categories for `WN-N`. Individual members might differ in terms of POS frequency distribution, again biasing the evaluation. One regular phenomenon accountable for this is the *verbification* of nouns and adjectives, when a verbal form is constructed without adding any derivational markers. While these derivations might to some extent preserve the similarities between words (e.g. *e-mail.N*→*e-mail.V* is similar to *fax.N*→*fax.V*), many cases are less transparent and benefit from POS separation (e.g. the meaning shift in *air.N*→*air.V* is different from *water.N*→*water.V*). One exception is the verb subset of SimLex which turns out to have a particularly low POS ambiguity.

2.6 Outlook

In the beginning of this chapter, we have hypothesized that the conceptual gap between the targets of vector space models and the targets presented in common evaluation scenarios might affect the performance measurements. We have shown that this indeed takes place in two experiments: one based on standard similarity benchmarks, and one based on a novel word class suggestion schema. This finding opens several interesting directions for future research that we briefly outline below.

Lexical unit-based modeling. We have shown that lemmatization and subsequent POS disambiguation benefit both benchmark- and resource-based performance of word embeddings. While verb semantics is notoriously hard to pinpoint per se, we show that modeling verbs via type-based distributed representations introduces additional grammar-related challenges which can be partially addressed with lemmatization and POS-disambiguation of the inputs. From the conceptual perspective, lemmatized and POS-tagged targets can be seen as another step towards conceptually plausible lexical unit-based modeling of word usage. In this work, we focused on single-word entities, and analyzing the effect of including multi-word expressions into the VSM vocabulary is an important direction for future studies.

Word embedding methods. To ensure fair comparison and to keep our evaluation setup compact, we have consciously restricted the scope of the study to a single word embedding model (SGNS), single window-based context size (`w2`) and a single parameter set (*word2vecf* SGNS default). Our results could be further validated by experimenting with alternative context definitions and word embedding models, e.g. GloVe (Pennington et al., 2014) and CBOW (Mikolov et al., 2013b). Experiments on character-based type-level models, e.g. *fastText* (Bojanowski et al., 2017) or *Charagram* (Wieting et al., 2016) could be another interesting extension to our work. However, it is not clear how to integrate lemma and POS information into the character-based representations in an elegant way.

Cross-linguistic studies. POS tagging and lemmatization are general and well-defined language-independent operations. We have focused on English and have shown that POS-typing and lemmatization implicitly target several grammar-level issues in the context of word embeddings. Ebert et al. (2016) demonstrate that the improvements from lemmatization hold in a cross-lingual setup. While we believe our results to generally hold cross-linguistically, the relative contribution of POS disambiguation and lemmatization will inevitably depend on the typological properties of the language, constituting another topic for further research. The very recent release of the unified MultiSimLex dataset by Vulić et al. (2020) makes such a study possible.

Suggestion-based evaluation. The word class suggestion procedure has clear advantages: it is class-based, polysemy-aware, does not introduce additional complexity and does not require an annotated corpus for evaluation. It is resource-agnostic and only requires the target lexical resource to group words into classes. However, several drawbacks must be addressed before it can be used at large. Our leave-one-out scenario excludes *singleton classes*, i.e. classes that only have one member. This issue will become less severe with resource coverage increasing over time. The evaluation depends on resource and VSM vocabulary coverage, and for qualitative comparison between VSMs vocabulary intersection should always be taken into account. Alternative suggestion strategies might be explored, e.g. averaging among class members instead of selecting the closest prototype.

Application scenarios. We have introduced word class suggestion as an evaluation benchmark for word embeddings. However, the WCS output might be used in vocabulary-based application scenarios, e.g. as annotation study support in cases when the lexicon is available, but the usage corpora are scarce; as a lexicographer tool for finding the gaps in existing lexica; and as a source for context-independent unknown word class candidates in a word sense disambiguation setup.

Contextualized encoders. One of the greatest successes of NLP in the past years is the introduction of strong contextualized pre-trained encoders like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and Flair (Akbik et al., 2019). Despite having lost the state-of-the-art status, static word embeddings are still widely used both in NLP and in practical applications. Throughout this chapter we have solely focused on static, non-contextualized word embeddings, as lexical semantics is easier to conceptualize in the static setting. Contextualized encoders generate context-dependent representations of words and are capable of implicit word sense disambiguation (Peters et al., 2018), prompting the re-design of the traditional evaluation toolkits. An example of a successful adaptation to a contextualized setup is the Word in Context task (Pilehvar and Camacho-Collados, 2019) that measures the ability of the model to disambiguate words given a sentence they appear in. Adapting our hypothesis and formulating a research question in the contextualized setting is another interesting direction for follow-up work. New work by Vulić et al. (2020) suggests a range of protocols for evaluating contextualized encoders like BERT in the context-free similarity benchmark setting; in particular, they devise static representations for word types by averaging over a set of contextualized encodings extracted from a corpus. While they only manipulate the *size* of this "usage example" pool, filtering it based on lemma and POS instead could be a viable direction for expanding our experiments to the contextualized use case.

2.7 Chapter Summary

In this chapter we have explored the effects of conceptual alignment between vector space models and evaluation benchmarks on the benchmark performance. We have provided a historical overview of static word embedding models, existing evaluation methods and known weaknesses of the existing static embedding architectures.

We have designed two experiments to explore our research hypothesis. In the first experiment we systematically evaluated the well-known word2vec SGNS model in a range of settings that aim to conceptually bridge the vector space model and the evaluation benchmarks. We have defined criteria for benchmark selection and shown on two representative benchmark datasets that the hypothesized discrepancy between type-based VSM targets and lexical unit-based benchmarks indeed harms the measured performance and penalizes the pre-trained models.

To validate and elaborate on our findings, we have designed a novel evaluation procedure for static word representations – Word Class Suggestion – and performed the second experiment using two widely used lexical-semantic resources: WordNet and VerbNet. The results of the experiment have confirmed our previous hypothesis and allowed deeper insights into the magnitude of the observed effect and its dependency on the resource, part of speech and systematic ambiguity. A range of recommendations and future work directions conclude the chapter.

Chapter 3

Formalism Matters

Massive pre-trained contextualized word and sentence representations have caused a methodological shift in natural language processing. Probing aims to analyze these representations to understand what aspects of language are captured by the models during pre-training. Linguistics is a convenient theoretical framework for this. More often than not, linguistics offers several ways to describe the same phenomenon, and the visibility of linguistic theories in the NLP landscape is largely determined by the availability of the corresponding lexical resources and corpora. Any linguistics-based probing study thereby commits to the formalisms used to annotate its data. But would the findings hold if another linguistic formalism were used instead? Can the choice of linguistic theory affect the probing results? We investigate this question using role semantics as a prominent multi-formalism phenomenon, and find that the formalism indeed matters.

3.1 Introduction

The emergence of deep pre-trained contextualized encoders has had a major impact on the field of natural language processing. Boosted by the availability of general-purpose frameworks like AllenNLP (Gardner et al., 2018) and Transformers (Wolf et al., 2019), pre-trained models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have caused a shift towards simple architectures where a strong pre-trained encoder is paired with a shallow downstream model, often outperforming the intricate task-specific architectures of the past.

The versatility of pre-trained representations implies that they encode some aspects of general linguistic knowledge (Reif et al., 2019). Indeed, even an informal inspection of layer-wise intra-sentence similarities (Figure 3.1) suggests that these models capture elements of linguistic structure, and those differ depending on the layer of the model. A grounded investigation of these regularities allows interpretation of the model's behavior, design of better pre-trained encoders and can inform downstream model development. Such investigation

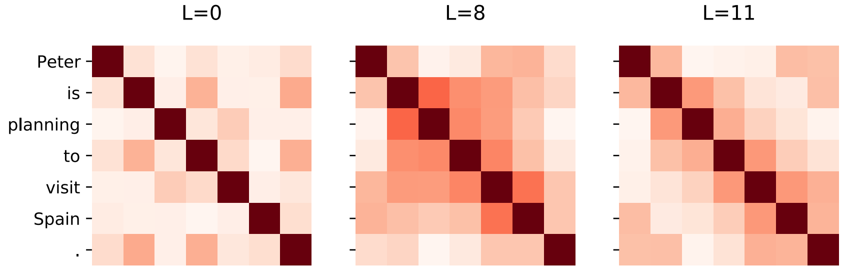


Figure 3.1: Word-wise intra-sentence similarity by layer L of the multilingual BERT-base. Functional words are similar in $L = 0$, syntactic groups emerge at higher levels.

is the main subject of probing, and recent studies confirm that BERT implicitly captures many aspects of language use, lexical semantics and grammar (Rogers et al., 2020).

3.1.1 Motivation

Most probing studies use linguistics as a theoretical scaffolding and operate on the task level. However, there often exist multiple ways to represent the same linguistic task: for example, English dependency syntax can be encoded using a variety of *formalisms*, incl. Universal (Schuster and Manning, 2016), Stanford (de Marneffe and Manning, 2008) and CoNLL-2009 dependencies (Hajič et al., 2009), all using different label sets and syntactic head attachment rules. Any probing study inevitably commits to the specific theoretical framework used to produce the underlying data. The differences between linguistic formalisms, however, can be substantial.

Can these differences affect the probing results? This question is intriguing for several reasons. Linguistic formalisms are well-documented, and if the choice of formalism indeed has an effect on probing, cross-formalism comparison will yield new insights into the linguistic knowledge obtained by contextualized encoders during pre-training. If, alternatively, the probing results remain stable despite substantial differences between formalisms, this prompts further scrutiny of what the pre-trained encoders in fact encode. Finally, on the reverse side, cross-formalism probing might be used as a tool to empirically compare the formalisms and their language-specific implementations. To the best of our knowledge we are the first to explicitly address the influence of formalism on probing.

Ideally, the task chosen for a cross-formalism study should be encoded in multiple formalisms using the same textual data to rule out the influence of the domain and text type. While many linguistic corpora contain several layers of linguistic information, having the same textual data annotated with multiple formalisms for the *same* task is rare. We focus on role semantics – a family of shallow semantic formalisms at the interface between syntax and propositional semantics that assign roles to the participants of natural language utterances, determining *who* did *what* to *whom*, *where*, *when* etc. Decades of research in theoretical linguistics have produced a range of role-semantic frameworks that have been operationalized in NLP: syntax-driven PropBank (Palmer et al., 2005a), coarse-grained VerbNet (Schuler, 2005), fine-grained FrameNet (Baker et al., 1998), and, recently, decompositional Semantic Proto-Roles (SPR) (Reisinger et al., 2015b; White et al., 2016b). The SemLink project (Bonial et al., 2013) offers parallel annotation for PropBank, VerbNet and FrameNet for English. This allows us to isolate the object of our study: apart from the role-semantic labels, the underlying data and conditions for the three formalisms are identical. SR3de (Mújdricza-Maydt et al., 2016) provides compatible annotation in three formalisms for German, enabling cross-lingual validation of our results. Combined, these factors make role semantics an ideal target for a cross-formalism probing study.

A solid body of evidence suggests that encoders like BERT capture syntactic and lexical-semantic properties, but only few studies have considered probing for predicate-level semantics (Tenney et al., 2019b; Kovaleva et al., 2019). To the best of our knowledge, we are the first to conduct a cross-formalism probing study on role semantics, thereby contributing to the line of research on how and whether pre-trained BERT encodes higher-level semantic phenomena.

3.1.2 Contributions

- We conduct cross-formalism experiments on PropBank, VerbNet and FrameNet role prediction in English and German, and show that the *formalism can affect probing results* in a linguistically meaningful way. We demonstrate that layer probing can detect subtle differences between implementations of the same formalism in different corpora and languages.
- On the technical side, we advance the recently introduced edge and layer probing framework (Tenney et al., 2019b); in particular, we introduce *anchor tasks* – an analytical tool inspired by feature-based systems that allows deeper qualitative insights into the pre-trained models' behavior.
- In a range of additional experiments, we explore *the effects of data size, dataset choice and task formulation* on layer probing results. We show that while extreme low-resource settings can prevent the probe from learning the scalar mix, a small amount of data might suffice for a general intuition about the layer utilization of a particular task. On a subset of Universal Dependency and POS probes, we show that the choice of a dataset does not seem to affect the layer probing results in a principled way. Finally, we highlight the caveats of using layer probes for sentence-level phenomena.

3.2 Contextualized representations

This section introduces a multitude of concepts and ideas that guide our study. We start with the discussion of ambiguity and contextualization, and review the main approaches to contextualization pre-dating the pre-trained encoder era. We then turn to the discussion of strong modern pre-trained encoders and their properties and introduce the BERT model which is the main object of our probing study.

3.2.1 Ambiguity and Context

We have already confronted ambiguity in our discussion of the conceptual gap between text and lexicon in Chapter 2, and now proceed to a more in-depth discussion of ambiguity and context in language. Natural languages maintain a delicate balance between the effort of the speaker, who aims to be concise, and the effort of the listener, who aims to understand the message perfectly. Ambiguity and context play a key role in achieving that balance. As an illustration from engineering, one reason programming is hard to learn is that programming languages are precise. The lack of ambiguity allows efficient and straightforward interpretation, but demands an unusual level of detail and clarity from the human "speaker" who is accustomed to natural languages. Any programmer who has once switched from a statically typed language like Java to a dynamically typed language like Python can both attest to the everyday convenience of a little bit of ambiguity, and experience the associated risks while working with another person's code. Unlike formal languages, natural languages are highly ambiguous and use explicit and implicit context and world knowledge to compensate for that. Words, sentences and texts are not meant to exist in isolation, and any NLP model that fails to take this into account would be severely limited in its predictive capabilities. Ambiguity transcends all language levels, and below we review some of the most important types of natural language ambiguity.

Morphological ambiguity. Due to the syncretism of natural language grammar, multiple morphological forms of a word can conflate into a single word type. Examples are abundant. English *run* can be interpreted as 1st or 2nd person singular (*I/you run*) or plural (*We/you/they run*), whereas the pronoun *you* itself can be singular or plural; the Russian *stol* (*table*) can be Nominative or Accusative; the German *kaufen* (*buy*) can be both infinitive (*I will ein Auto kaufen*) and 1nd or 3nd person plural (*Wir/sie kaufen ein Auto*). This phenomenon is omnipresent and rarely causes any issues to language speakers, who can derive the correct interpretation from the sentential context. A particular case of morphological ambiguity is part-of-speech ambiguity: a *saw* can be both past tense of *see* as a verb, and singular of *saw* as a noun.

Lexical ambiguity. Discussed in detail in Chapter 2, lexical ambiguity is a common phenomenon in natural language: *bat.N* can refer to an animal or sports inventory item; *April.N* can refer to month or to a person's name; *get.V* might be synonymous to *obtain.V* as in *"He got a car"* or to *become.V* as in *"He got sick"*. Sense disambiguation happens in sentential context. However, discourse-level context might suffice as well, leading to competitive performance of the most frequent sense baselines as measured on sense-annotated corpora: intuitively, if we know that the text is about sports, encountering a *bat.N* in the *animal* sense becomes less likely.

Syntactic ambiguity. Even with morphological and lexical ambiguity resolved, sentence structure poses additional challenges to the interpretation. One prominent example of syntactic ambiguity is prepositional phrase attachment. Consider three sentences: (1) *"I see a star with a telescope"*, (2) *"I see a dog with a telescope"* and (3) *"I see a man with a telescope"*. Does the phrase *"with a telescope"* relate to the verb *"see"* or to the observed object? Based on purely grammatical analysis, both interpretations are valid; however, a language speaker would easily disambiguate (1) and (2) based on world knowledge, while (3) would either be accepted as truly ambiguous or require wider discourse context to resolve.

Proposition-level ambiguity. The resolution of ambiguity at syntactic level leaves us with proposition-level ambiguity that also needs to be resolved. One prominent example of proposition-level ambiguity is semantic role assignment. In the sentences (1) *"A man hit the table"* and (2) *"A hammer hit the table"*, *"man"* and *"hammer"* are both grammatical subjects, however, they have different semantic roles: `Agent` in (1), and `Instrument` in (2). Resolving this syntax-level ambiguity is impossible without external lexical knowledge and context.

Most NLP applications target word sequences and not merely isolated words: named entity recognition aims to label multi-word text spans, relation extraction induces relations between entities, question answering needs to interpret the question (usually, a sentence) and provide an answer, and machine translation aims to convert sentences in one language to another. All these tasks require some form of sequence-level interpretation, which is impossible without resolving natural language ambiguity on each of the levels – explicitly, as in traditional pipeline-based NLP, or implicitly, as in state-of-the-art end-to-end neural encoders.

3.2.2 Context as feature

Even the earliest NLP systems incorporated contextual information into the decision-making. Since sophisticated context representations were not available at the time, this has first taken the form of explicit context patterns in template-based and rule-based systems (Chinchor et al., 1993). The use of explicit contextual patterns, however, led to poor generalization, low recall, and required extensive expert involvement at the development stage.

More flexibility could be gained through feature-based context representations that served as input to machine learning classifiers; for example, to assign a role to a semantic argument, a feature-based semantic role labeling system would often consider the syntactic path between the argument and the predicate, the head word of the predicate, the relative position of the predicate with respect to the argument token (left or right), the preposition lemma for the arguments that are prepositional phrases, etc. (Xue and Palmer, 2004; Björkelund et al., 2009). To combat sparsity, linguistic generalizations were incorporated into the feature sets, e.g. the feature "word to the right" would be supplemented by the feature "POS of the word to the right". Distributional representations have been used to further reduce sparsity and enrich the models with type-level lexical information: for example, Roth and Woodsend (2014) incorporate static predicate and argument embeddings into the feature set of a standard semantic role labeler and report improved performance both in- and out-of-domain.

Another way to reflect the sequential nature of the NLP tasks is to use models specifically designed for representing sequential phenomena. Hidden Markov Models have been successfully used for low-level tasks that do not require large context to resolve ambiguity like part-of-speech tagging (Charniak et al., 1999), however, they suffered from the limitations of the Markovian assumption and failed to account for long-range phenomena. To model global sequence structure, Conditional Random Fields (Lafferty et al., 2001) have been successfully employed for the tasks with strong interdependencies between labels like POS tagging (Lafferty et al., 2001), Named Entity Recognition (Settles, 2004), parsing (Finkel et al., 2008) and semantic role labeling (Toutanova et al., 2008).

Models based on discrete linguistic features had a multitude of advantages, incl. better interpretability and the ease of ablation experiments, but suffered from low efficiency, generalization issues and high development costs. Since most feature-based models required linguistic input to aid generalization, they were embedded in NLP pipelines. This led to error propagation and harmed reproducibility as the pipeline components were subject to change: for example, the improved performance of an SRL system could in reality stem from a better upstream parser and not from the improvements to the SRL module itself (He et al., 2018). The reliance on the pipeline required the training data to contain gold annotations not only for the target phenomenon, but also for all the upstream components, making the creation of training data expensive and limiting the NLP system development to a few richly annotated corpora like Penn Treebank (Marcus et al., 1993) and OntoNotes (Hovy et al., 2006). Managing models and components presented a significant development overhead, and although meta-frameworks for customizing pipelines were available (Cunningham et al., 2002; Loper and Bird, 2002; Eckart de Castilho and Gurevych, 2014), the prospect of efficient one-step learning from and labeling of raw text remained attractive, motivating the development of end-to-end NLP systems.

In the context of our work, we highlight that the superior performance and convenience of today's end-to-end models compared to their feature-based counterparts does not diminish the value of expert-curated feature sets and system architectures: the reason why feature-based systems have become obsolete is not the linguistic modeling choices, but the practicalities of building a good, robust, expressive target and context representation. We believe that creative re-use of the ideas accumulated during the feature-based NLP era is a promising and underrepresented topic; our anchor-based probing methodology described later in this chapter is a step in this direction.

3.2.3 End-to-end context modeling

The previously discussed work of Collobert and Weston (2008) has shown that it is possible to achieve near state-of-the-art performance on a range of NLP tasks without relying on pipeline-based architectures and explicit linguistic features. Their neural end-to-end architecture used convolutional neural networks to aggregate information from the context; however, the key to the success of their system was the use of pre-trained word representations obtained from a large unlabeled corpus.

Over the next years, most NLP tasks have transitioned to the end-to-end neural setting, showing comparable or better performance and far better efficiency than their feature-based predecessors. While Collobert and Weston (2008) used convolution over a window of words to represent context, a more robust solution for representing sequential information was found in bidirectional long-short term memory networks (BiLSTM, Hochreiter and Schmidhuber (1997)) – a variant of recursive neural network more suited for representing long-distance dependencies in the sentence. Modeling sub-word information was found to increase model generalization, and convolutional neural network (CNN)-based representation over word characters became a popular way to allow for more flexibility in the end-to-end models. As before, conditional random field should be used to account for the global tag structure and constraints, this time with all-neural inputs produced directly from raw text. The resulting BiLSTM-CNN-CRF architecture (Ma and Hovy, 2016) has become a cornerstone of end-to-end NLP and has been applied to the majority of NLP tasks with great success.

One important structural weakness of recursive neural networks is their inability to handle long-distance dependencies between words due to vanishing gradients. This hampers the ability of vanilla RNNs to model NLP tasks that require long-distance dependencies to be taken into account, e.g. question answering and semantic role labeling. BiLSTMs have partially addressed this problem by enhancing RNNs with a memory mechanism, however, the problem of vanishing gradients still persisted in deep BiLSTM models. One creative way to overcome this limitation was proposed by Marcheggiani and Titov (2017) who use a combination of deep BiLSTM with a graph convolutional network over the syntactic trees, allowing the model to “teleport” to linearly distant, but syntactically close words, improving the semantic role labeling performance on long-range dependencies. This, however, rendered their system dependent on syntactic preprocessing.

A more principled solution to modeling distant contexts was offered by attention-based models: unlike RNNs and BiLSTMs which are inherently sequential, attention models take *full* context as input and *learn* the relative importance of the context items during training. Attention has successfully replaced sequential models for a variety of NLP tasks that require handling of long inputs like parsing (Dozat and Manning, 2016) and semantic role labeling (Strubell et al., 2018). A variant of attention – self-attention – is a key component of the Transformer model that powers most of the modern pre-trained contextualized encoders.

End-to-end models have been shown to greatly benefit from incorporating non-contextualized, static word representations along with task-specific representations learned during training. However, the contextualized representation itself still had to be learned based on the labeled data for a given end-task, and such labeled data was scarce. The ability to outsource the learning of contextual representations to the pre-training stage was crucial to further NLP progress, leading to the development of strong pre-trained contextualized encoders.

3.2.4 Strong pre-trained encoders

Just as static pre-trained word embeddings can be seen as a lookup function that given a *word type* returns its vector representation, a dynamic contextualized encoder is a function that given a *token* and its surrounding context (usually, a sentence) produces a contextualized vector representation for that token as output. Operating on token instead of type-level allows greater flexibility due to implicit modeling of word sense ambiguity and other context-sensitive phenomena.

Early evidence of the benefits of pre-trained sequential modeling is presented in *context2vec* (Melamud et al., 2016) who suggested using BiLSTM context encoding instead of naive window-based word vector averaging and demonstrated improved lexical substitution and word sense disambiguation performance. McCann et al. (2017) propose CoVe, a context-sensitive deep BiLSTM encoder trained on the machine translation task, and show that when used as a drop-in replacement for static GloVe embeddings, it improves the performance on a variety of classification and question answering tasks. ELMo (Peters et al., 2018) and Flair (Akbik et al., 2019) are BiLSTM-based context embedders trained with a language modeling objective that allows pre-training on large unlabeled monolingual corpora. GPT (Radford et al., 2019) and BERT (Devlin et al., 2019) are Transformer-based models (Vaswani et al., 2017) trained with a language modeling objective; the derivatives of these models power the current state of the art for the majority of NLP tasks. Our study is based on a particular BERT model – multilingual mBERT, – which we describe in detail later in this chapter; below we outline some key properties that are important for our further discussion and hold for modern pre-trained contextualized encoders in general.

Subword inputs. Most modern contextualized encoders utilize some form of subword modeling to target the out-of-vocabulary issues and allow sharing of information between word forms. ELMo employs a character CNN as the input representation; Flair is a purely character-based model devoid of the notion of a token: sequences are encoded on character level, and the representation is extracted from the characters at the token boundaries at inference time. GPT-2 and BERT use variations of byte-pair encoding to create a subword vocabulary that is later used to tokenize the raw inputs.

Layered structure. The advantage of deep neural network architectures over shallow ones has been established early in the work on end-to-end NLP. Modern pre-trained contextualized encoders usually consist of several deep layers; allowing the end-model to utilize information from several layers leads to a better end-task performance as the layers appear to reflect different levels of abstraction over the input sequence (Peters et al., 2018; Devlin et al., 2019). The difference in layer utility depending on the end task is the key component of layer probing, which is used as a core framework in our work and is discussed in detail below.

Fine-tuning. A pre-trained contextualized encoder might be used in two principled ways: as initialization for the task-specific model (in which case the pre-trained weights are fine-tuned during task training) and as a feature generator (in which case the encoder model is not updated, i.e. *frozen*). Both modes have their use-cases: best performance is often achieved when the training signal is propagated throughout the encoder; however, this implicitly increases the power of the end-task model and requires significant training time. Using pre-trained encoders as feature generators often results in acceptable performance at low computational cost; the architectural choices of state-of-the-art models like BERT make it easy to adapt them to any NLP task by simply adding a classification layer on top of the encoder (Wolf et al., 2019).

Stacking. As with task-specific end-to-end models, combining the outputs of different pre-trained contextualized encoders often results in superior performance, as the representations learned by contextualized and static word embedding methods are complementary. For example, Akbik et al. (2019) advocate for combining contextualized flair embeddings with static GloVe embeddings as input and demonstrate the superiority of this approach on NER; while Peters et al. (2018) obtain the same effect by concatenating GloVe embeddings with ELMo encoder outputs, although to a smaller extent.

3.2.5 Post-BERT models and Open Challenges

Since this area of study develops quickly, providing a comprehensive and up-to-date snapshot of research in Transformer-based contextualized encoders is a separate challenge and lies outside the scope of our work. We thereby focus on a selection of recent developments relevant to probing, and refer to a recent survey by Xia et al. (2020) for a comprehensive overview of modifications to the training objective and regime, pre-training data, attention mechanism, etc.

New languages and domains. The original BERT model was based on English. Since both subword tokenization and the language modeling objective in BERT are unsupervised, it is possible to train language-specific BERT models, e.g. CamemBERT for French (Martin et al., 2020), BERTje for Dutch (de Vries et al., 2019) and RoBERT for Romanian (Masala et al., 2020). A separate line of research is dedicated to multilingual models: Devlin et al. (2019) have released mBERT, accompanied by XLM (Conneau and Lample, 2019) and followed by XLM-RoBERTa (Conneau et al., 2020a) and others, all demonstrating impressive multi-lingual transfer capabilities. Specialized models have been developed for data-rich domains, e.g. SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2019). Pre-training BERT encoders for underresourced languages and domains remains an open challenge, and the factors that enable multilingual learning are subject to active investigation. Our probing studies deliver new insights on the layer utilization in the pre-trained multilingual BERT model.

Model size and efficiency. Multiple works have demonstrated that the performance of the pre-trained contextualized encoders scales with the model and dataset size (Conneau et al., 2020a); top-performing Transformer-based contextualized models might have millions (billions) of parameters, which hampers their deployment and slows down end-task fine-tuning. It has been observed that most of these models are in fact overparametrized (Kovaleva et al., 2019; Michel et al., 2019; Prasanna et al., 2020). Several parallel research efforts aim at reducing the computational and environmental (Strubell et al., 2019) footprint via model distillation (Sanh et al., 2020), pruning (Michel et al., 2019) and adjusting the computationally expensive self-attention mechanism (Sukhbaatar et al., 2019). A related line of research in adapters (Houlsby et al., 2019; Pfeiffer et al., 2020) replaces fine-tuning the full model with injecting trainable weights into the layers of an otherwise *frozen* pre-trained encoder, demonstrating competitive performance at a much lower computational cost. Our work contributes to the better understanding of the functions of mBERT's layers.

Sentence representations. One of the key features of the original BERT architecture is its ability to represent both words, sentences and sentence pairs in a unified fashion thanks to a combined word- and sentence-level training objective. Subsequent work has demonstrated that sentence representations produced by the original BERT model are far from perfect (Reimers and Gurevych, 2019). Numerous studies have focused on obtaining better sentence-level models based on BERT, ranging from fine-tuning on semantic textual similarity and natural language inference data (Reimers and Gurevych,

2019) to a creative use of discourse context in a self-supervised setting (Nie et al., 2019). A related line of work on document-level representations exploits domain-specific discourse context to fine-tune a pre-trained SciBERT model using citation graphs in the scientific domain (Cohan et al., 2020). Our study delivers insights on sentence-level representations produced by pre-trained mBERT.

Evaluation. Pre-trained contextualized encoders are traditionally evaluated using benchmark kits like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) that encompass a large variety of language understanding tasks, including semantic textual similarity (Cer et al., 2017), natural language inference (Conneau et al., 2018b), paraphrasing (Dolan and Brockett, 2005), word similarity in context (Pilehvar and Camacho-Collados, 2019) and others. Widely used evaluation datasets are known to contain annotation artifacts that contextualized encoders readily exploit (Gururangan et al., 2018). This leads to inflated performance estimates and vulnerability of the trained models to adversarial attacks[1]. Devising more robust annotation (Bowman et al., 2020) and evaluation protocols (Ribeiro et al., 2020) and creating more challenging and robust evaluation datasets (Sakaguchi et al., 2020; Nie et al., 2020) is an active line of research closely related to behavioral probing, as discussed below.

Interpretability and Bias. The increase in performance of the pre-trained contextualized encoders is accompanied by the increasing complexity of the models. The lack of control over pre-training data sources (Xia et al., 2020), bias (Kurita et al., 2019; Tan and Celis, 2019) and susceptibility to adversarial attacks motivate the research in interpretable and explainable models. State-of-the-art approaches to model interpretation include detecting the tokens that contribute to the prediction via saliency maps (Simonyan et al., 2014) as well as input manipulation via token replacement (Ebrahimi et al., 2018) and input reduction (Feng et al., 2018); a separate line of work investigates explicit explanation generation at prediction time (Camburu et al., 2018). Several tools have been proposed to make model interpretation more accessible (Vig, 2019; Wallace et al., 2019); however, the goal of explainable deep NLP is far from being reached as the community debates the validity and utility of the existing approaches to model interpretation (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Atanasova et al., 2020). Research in interpretability shares a lot of methodological ground with probing.

3.2.6 The (m)BERT model

The multilingual BERT model (mBERT) used in this work is the classic BERT architecture trained on a 104-language corpus sourced from Wikipedia. Although a multitude of better-performing encoder models have been proposed

[1] The general issue of shortcut learning is not new and not specific to NLP, see e.g. (Geirhos et al., 2020) for a high-level overview of the related issues in computer vision, medical imaging, etc.

since BERT's release in 2018 (Liu et al., 2019; Lan et al., 2019; Raffel et al., 2019), we chose the original BERT model for several reasons. First, similar to static word embeddings, current evaluation methodology conflates the merits of the model with the performance of a model *instance* which could have been trained on more data or have higher capacity; occupying a higher position in GLUE or SuperGLUE thereby does not signal conceptual superiority of a model per se. Second, most of the research in probing is focused on BERT[2]; using the same model architecture allows us to inherit the probing methodology from related work and contextualize our findings. Although this focus on a single family of models has been recently criticized as over-investment (Xia et al., 2020), we believe that knowing few models well is more desirable than scattering the community effort to keep up with state of the art. In addition, many of the analysis techniques refined on BERT can be transferred to other, more advanced transformer-based models; for example, the layer probing methodology used in our study only requires the encoder to associate input tokens with a layer-wise token representation with the information propagated through the layers in one direction.

Architecturally, mBERT is identical to the original BERT model which we now briefly review. Introduced in (Devlin et al., 2019), BERT is a bi-directional contextualized Transformer encoder that uses masked language modeling (MLM) and next sentence prediction (NSP) pre-training objectives to learn sequence representations from raw, unlabeled textual data. Unlike the earlier ELMo model (Peters et al., 2018), which works around bidirectionality by concatenating the outputs of separately trained forward and backward LSTM models, BERT is deeply bidirectional in the sense that the forward and backward language models are trained jointly. Since the traditional next word prediction objective would allow the information about a word to "leak" through context, BERT employs the MLM objective which resembles the Cloze task: a portion of input tokens is replaced with a special *[MASK]* token at random, and the model is tasked with predicting the original token given the surrounding sequence. In addition BERT employs the second, document-level next sentence prediction objective which is formulated as a binary task given two input sentences.

The BERT model closely follows the Transformer encoder architecture presented in (Vaswani et al., 2017). In the default configuration, BERT takes a sequence of WordPiece (Wu et al., 2016) subword tokens $s = [w_1, w_2 ... w_k]$ as textual input. This sequence is processed by a stack of architecturally identical Transformer encoder blocks $L = [L_1, L_2 ... L_m]$. Each encoder block applies multi-head self-attention to the input representation of each word in the sequence, and the updated word representation is propagated to the upper block through a feed-forward layer where it again serves as input. We further denote the input representation of the wordpiece i at layer n as w_i^n. After the final encoder block, a softmax layer is applied to project the model output into the pre-training objective tag space. The number of encoder blocks L, their hid-

[2] To the point where the NLP community has coined a dedicated name for this line of research: *Bertology* (Rogers et al., 2020)

den size H and the number of attention heads A are hyperparameters; the two widely used pre-trained instances of the BERT model are BERT-base ($L = 12$, $H = 768$, $A = 12$) and BERT-large ($L = 24$, $H = 1024$, $A = 16$).

Unlike sequential models, self-attention mechanism per se does not model word order; for this reason, the Transformer incorporates a consistent, dynamically generated positional embedding into the input representation. The BERT model supplements this with the segment embedding which denotes the part of the structured input a token belongs to (see below). The input sequence at L_1 is thereby encoded by summing up the randomly initialized wordpiece embeddings, the positional and the segment embeddings:

$$w_i^1 = wp(w_i) + pos(i) + seg(w_i)$$

Note that the encoding at L_1 is thereby context-agnostic as no self-attention update has been yet applied to the token representation; however, it already contains the positional information.

The result of BERT encoding is a layer-wise representation of the input wordpiece tokens with higher layers representing higher-level abstractions over the input sequence. A key advantage of the BERT model is its ability to jointly encode single tokens, sentences and sentence pairs via a structured input mechanism. This is achieved by introducing the special token *[CLS]* that is appended to each input sequence and serves as a representation for sentence and sentence pair classification tasks. A special *[SEP]* token is used to separate sentences in a sentence pair setup.

The implementation of BERT is very practical and allows an easy adaptation to the majority of downstream NLP tasks by replacing the final softmax over vocabulary with a shallow task-specific tag projection layer. As with the majority of the modern encoders, the BERT model can be fine-tuned or used as a feature generator. While best results are obtained through fine-tuning, the original work demonstrates that even the frozen model can yield competitive results. Since the embedding and attention weights are not updated when using the frozen model, the choice of the *layer* and the layer combination strategy become important (in line with the observations made for ELMo by Peters et al. (2018)). The original paper reports experiments on different layer combinations for the NER task, with the 4% F1 difference between the best (last four layers concatenated) and the worst (the wordpiece embedding layer L_1) configurations. Since layer utility is task-specific, a flexible solution introduced by Peters et al. (2018) is to use *scalar mix* – a weighted sum of the layers, with weights being learned jointly with the tag projection layer. This technique is a cornerstone of the layer probing approach introduced in detail in Section 3.5.

3.3 Probing

Having reviewed the core properties and challenges associated with the use and development of strong pre-trained contextualized encoders, we now turn to probing – the methodological backbone of our study. We briefly review the motivation and core approaches to probing and summarize what is known about the linguistic abilities of BERT and mBERT.

3.3.1 Why probe

Deep neural models have been responsible for a major increase in end-task NLP performance over the past decade. Unlike their feature-based predecessors, deep neural models lack transparency, making analysis and attribution of the models' predictions challenging. While boosting NLP performance is possible without understanding the internal mechanisms that drive this performance, gaining insights into the inner workings of the NLP models is worth pursuing for a range of reasons (Belinkov and Glass, 2019). From a *model development* perspective, it presents a viable alternative to the "shot-in-the-dark" approach to natural language processing, where the superiority of a particular model cannot be clearly attributed to the modelling choices. Along with strict control of evaluation setups, knowing where current models excel and where they fall short would contribute to iterative, hypothesis-driven research towards better representations. From an *application* perspective, transparency is key for safe, ethical and accountable AI. Finally, from a *linguistics* perspective, transparent NLP models would allow human interpretation of the representations learned by NLP models during training and help us better understand the reasons behind their practical power. The recent shift towards unified architectures based on shared pre-trained contextualized encoders creates an excellent environment for focusing the interpretability efforts on few commonly used models, and a lot has been discovered about the internal mechanisms of BERT in the few years since its release.

An important question in achieving transparency is the choice of the *framework* in terms of which the models are to be interpreted. While NLP models can be evaluated and – to some extent (Wallace et al., 2019) – interpreted using practical end-tasks like community-based question answering, citation context prediction or clinical trial report parsing, such an evaluation is limited to a particular application, conflates a range of complex phenomena required to solve the task, and is tied to the idiosyncrasies of the corresponding application domains. Strong pre-trained encoders have made it possible to construct efficient and seemingly transparent few-shot models that allow the user to trace the prediction back to the training data instance (e.g. Yang and Katiyar (2020)). However, this approach doesn't explain *why* a certain training instance has high similarity to the input and does not generalize to other tasks; besides, not every end task can be cast as a zero-shot learning problem.

An alternative approach to model analysis is probing (Conneau et al., 2018a). Instead of end tasks, probing focuses on basic competences of the model, from counting words in the input sentence to uncovering latent predicate-semantic representations and co-reference. Several factors make linguistics an attractive interpretation framework: linguistic tasks are well-studied and well defined; linguistic tasks are grounded in theory and designed to generalize across

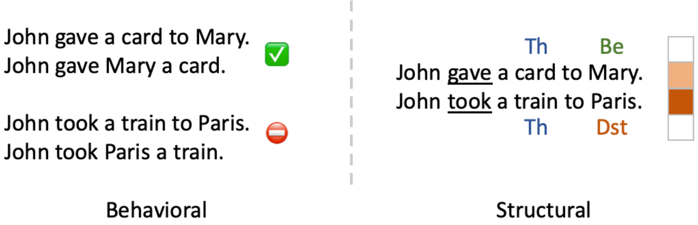

John gave a card to Mary.
John gave Mary a card.
✓
John took a train to Paris.
John took Paris a train.
⛔
Behavioral
Th
Be
John gave a card to Mary.
John took a train to Paris.
Th
Dst
Structural

Figure 3.2: Behavioral and structural probing for role semantics; while a behavioral probe would focus on mesasuring the models capability to detect ungrammatical constructions, a structural probe would focus on finding the components of the model that are used to make role distinctions.

languages and domains; finally, linguistic tasks are often backed by expert-annotated corpora, and two decades of feature-based NLP provide valuable insights into the modelling of the underlying phenomena.

3.3.2 Probing Methodology

Following an analogy from experimental psychology, probing approaches can be roughly divided into two groups: behavioral and structural probing (Belinkov et al., 2020), the former focusing on measuring the model's linguistic competence, the latter analysing the internal mechanisms of the model with respect to linguistic properties. The two lines of research are complementary and we briefly discuss them below; Figure 3.2 illustrates the differences between behavioral and structural probing on a high level.

Behavioral probing measures the models' linguistic abilities by constructing challenge sets focused on specific linguistic phenomena. Common examples of behavioral test sets include the GLUE and SuperGLUE benchmarks (Wang et al., 2018, 2019). Poliak et al. (2018) propose a set of natural language inference tasks annotated with diverse linguistic phenomena like factuality, puns, gendered anaphora and sentiment; and Warstadt and Bowman (2020) annotate the linguistic acceptability corpus CoLA (Warstadt et al., 2019) to evaluate pre-trained models against a range of fine-grained syntactic phenomena. Methodologically, approaches to behavioral testing of pre-trained models also include input manipulation (Balasubramanian et al., 2020) and masked behavioral probes that are cast as a fill-in-the-blank task similar to BERT's pre-training objective (Petroni et al., 2019).

The main advantages of the behavioral approach for probing is the control over the evaluation setup, the ability to account for rare, long-tail phenomena and to employ creative input manipulation techniques and synthetic data. Among the drawbacks of behavioral probing are the lack of explanations behind the models' performance: if the performance is low, it is unclear which aspect of the model is responsible for it; if the performance is high, it does not guarantee that the model will perform well in the end-task scenarios where the distribution of the phenomena might be different (similar to the points made for static word embeddings by Faruqui et al. (2016)). Constructing a good behavioral benchmark is another challenge: a too small benchmark might be not representative and easily overfit; a large, high-coverage benchmark is costly to produce, and crowdsourcing such datasets results in biases and readily exploited artifacts (Gururangan et al., 2018) that are not easy to mitigate (Bowman et al., 2020). Contemporary work in template-based evaluation kits inspired by software unit tests (Ribeiro et al., 2020)[3] offers a promising solution to finding the balance between benchmark coverage, accounting for long-tail phenomena and controlling for bias.

Structural probing focuses on the encoding of linguistic phenomena by different components of a pre-trained model. Belinkov and Glass (2019) organize the structural probing approaches by the method, the linguistic target and the model component. The mainstream approach to structural probing is to design and train a classifier model that uses the frozen pre-trained encoder representations as input. Since the encoder is not updated and the classifier is simple, high performance on the task is interpreted as the necessary information being *easily extractable* from the pre-trained model. Structural probing commonly focuses on classic linguistic targets like part of speech, syntactic relations (Hewitt and Manning, 2019), coreference and semantic roles (Tenney et al., 2019a; Kovaleva et al., 2019), often sourced from the same richly annotated corpus. Recent work suggests a structural probing methodology to study the encoding of lexical-semantic information (Vulić et al., 2020). From the model component side, the two most common analysis targets for Transformer models are the encoder layers (Peters et al., 2018; Devlin et al., 2019; Tenney et al., 2019b; de Vries et al., 2020) and the self-attention heads (Kovaleva et al., 2019; Hewitt and Manning, 2019; Kulmizev et al., 2020; Voita et al., 2019). Compared to the behavioral approach, structural probes offer deep insights into the internal mechanics of the pre-trained encoders, and studies in structural probing contribute to research on transparency, model size and overparametrization (Voita et al., 2019; Kovaleva et al., 2019) and multilinguality (Chi et al., 2020; de Vries et al., 2020), among others. Concurrent work in amnesic probing (Elazar et al., 2021) bridges the gap between behavioral and structural analysis by determining the components of the model responsible for representing a certain linguistic phenomenon, ablating these components and measuring the impact on the end-task performance.

Structural probing is associated with two important caveats. First, as Hewitt and Liang (2019) demonstrate, an overly expressive probing classifier might be able to fit any signal independent of the actual capabilities of the underlying pre-trained model; however, if the probe expressiveness is restricted, the failure to localize a certain linguistic phenomenon within a model might be attributed to the probe not being complex enough to capture it. Recent work by Zhu and Rudzicz (2020) refines the criteria for probe selection based on information theory, and Voita and Titov (2020) propose a new class of information-theoretic probes that mitigate the selectivity issues; an alternative line of research advocates for comparing the representations directly, e.g.

[3] Best paper award at ACL 2020.

via representational similarity analysis (RSA, Kriegeskorte et al. (2008)).

Second, the fact that certain linguistic information is *readily available* within the pre-trained model does not automatically imply that this information is in fact *used* when solving the end tasks (Elazar et al., 2021; Kovaleva et al., 2019). Indeed, high-level tasks like NLI have been shown to be solvable by relying on simple heuristics, and it remains an open question whether the use of linguistic information needs to be enforced in such a case. We believe that the combination of refined behavioral tests (Ribeiro et al., 2020) and ablation-based probing analysis (Elazar et al., 2021) is a promising research direction that would shed light on this issue.

Our work focuses on structural probing. In particular, we extend the layer probing approach proposed in (Tenney et al., 2019b) and (Tenney et al., 2019a), while following the recommendations from Hewitt and Liang (2019) to enforce the probe selectivity, and perform extensive experiments to determine the impact of linguistic formalism on the measured structural properties of the pre-trained mBERT model. However, our core question – the effect of a linguistic formalism on probing results – equally applies to behavioral probing, and we leave the design of the corresponding challenge sets as a promising avenue for future work.

3.3.3 Linguistic abilities of BERT

We now summarize the findings related to the linguistic abilities of the pre-trained BERT models. Since our work is dedicated to structural probing, we put particular focus on this area, and only briefly mention some important results from behavioral probing. For a comprehensive overview of the BERT properties discovered up to date, we refer to the survey by Rogers et al. (2020).

Pre-trained BERT models have a range of proven linguistic capabilities. It has been shown that contextualized word representations produced by pre-trained BERT models are suited for zero-shot word sense disambiguation and outperform state-of-the-art WSD models in some settings (Reif et al., 2019; Wiedemann et al., 2019). A multitude of independent studies (Goldberg, 2019; Tenney et al., 2019b) found evidence of syntactic knowledge in the pre-trained models; Hewitt and Manning (2019) were able to learn a transformation to reconstruct dependency trees from raw BERT representations of sentences. Petroni et al. (2019) demonstrate that pre-trained BERT models contain factual knowledge using a fill-in-the-blank probe. The evidence of the high-level semantic capabilities of pre-trained BERT is inconclusive. Tenney et al. (2019a) show that the English PropBank semantics can be extracted from the encoder and follows syntax in the layer structure. However, out of all role-semantic formalisms PropBank is most closely tied to syntax, and the results on proto-role and relation probing do not follow the same pattern. Kovaleva et al. (2019) identify two attention heads in BERT responsible for FrameNet relations. However, they find that disabling them in a fine-tuning evaluation on the GLUE (Wang et al., 2018) benchmark does not result in decreased performance. Ettinger (2020) show that BERT struggles with making fine-grained role-semantic distinctions but has a good grasp of hypernym prediction, and Warstadt and Bowman (2020) report that BERT models do not perform well on grammaticality judgements for rare constructions. Despite the impressive performance on the NER end-task, pre-trained BERT has been found brittle to input manipulation through named entity swapping (Balasubramanian et al., 2020).

Two main analysis targets in the structural probing have so far been the attention heads and the layers of the pre-trained BERT model.

Attention heads. The study by Reif et al. (2019) provides evidence that attention heads encode syntactic information; Kovaleva et al. (2019) demonstrate that most attention heads fall into a limited set of patterns and can be pruned without a significant impact on the end-task performance (sometimes even resulting in increased performance); Voita et al. (2019) elaborate on these findings by identifying *confident* attention heads that only focus on a small range of tokens and proposing a pruning technique that allows removal of the majority of attention heads in pre-trained BERT without major performance degradation. Hewitt and Manning (2019) use attention head weights to reconstruct dependency trees from the latent representation, and Chi et al. (2020) expand this methodology to the multilingual case.

Layers. The importance of layer-specific information in deep contextualized encoders was acknowledged early on: Peters et al. (2018) demonstrate that it is crucial to allow the end-task model access to the information from all encoder layers in the feature-based setup; in a similar setting, Devlin et al. (2019) show that selecting a subset of model layers for feature-based predictions can substantially boost end-task performance. Several probing studies confirm that a weighted combination of layers or averaging over a subset of layers results in more accurate probing task predictions than focusing on a single layer (de Vries et al., 2020; Vulić et al., 2020). Regarding the localization of linguistic information, a general consensus in the probing literature is that earlier layers of the model encode positional and lexical information, while higher layers are responsible for high-level phenomena like syntax and semantic roles. In a large-scale probing study, Tenney et al. (2019a) demonstrate that the order of information processing in the encoder layers of the BERT models resembles a classic NLP pipeline with low-level surface analysis tasks (POS, parsing) followed by higher-level semantic tasks (SRL, coreference). Recent work in lexical-semantic probing (Vulić et al., 2020) reports that the best aggregated representations for lexical tasks can be retrieved by averaging the lower layers of the model, while Chi et al. (2020) show that syntactic structure can be best recovered from the middle-to-late layers of the model. BERT models come in different sizes, however, Tenney et al. (2019a) demonstrate that the task-specific layer utilization in BERT-base and BERT-large follows the same pattern and scales with the size of the model. The amnesic probing study by Elazar et al. (2021) suggests that the layer importance for a particular task changes depending on whether the probe involves a masking operation; while the probe used in our study does not use masking, this intriguing finding might shed some light on the discrepancies between the behavioral and structural probing results.

Our probing methodology builds upon the edge and layer probing framework. The encoding produced by a frozen BERT model can be seen as a layer-wise snapshot that reflects how the model has constructed the high-level abstractions. Tenney et al. (2019b) introduce the edge probing task design: a simple classifier is tasked with predicting a linguistic property given a pair of spans encoded using a frozen pre-trained model. Tenney et al. (2019a) use edge probing to analyse the layer utilization of a pre-trained BERT model via scalar mixing weights (Peters et al., 2018) learned during training. We revisit this framework in Section 3.5.

3.3.4 Linguistic abilities of multilingual BERT

As already briefly mentioned, the multilingual Transformer models pre-trained on mixed-language corpora demonstrate surprising effectiveness incl. zero-shot cross-lingual transfer capability for POS tagging and named entity recognition, even for the languages that use different scripts (Pires et al., 2019). Conneau et al. (2020a) show that a multilingual model can perform on-par with its monolingual counterpart and stress the importance of data size and cross-lingual dataset balancing. They coin the term *the curse of multilinguality*, highlighting the importance of increasing the model capacity with the number of languages. Conneau et al. (2020b) demonstrate that neither the shared vocabulary nor the joint pretraining are necessary for multilinguality and further show that contextualized representations for different languages can be aligned post-hoc. The results from a range of synthetic experiments by Dufter and Schütze (2020) suggest that the key factors contributing to the multilinguality of mBERT are the under-parametrization of the model (that forces it to learn multilingual abstractions), shared special tokens and position embeddings, and word order; they apply these intuitions to train a larger-scale model for Hindi, English and German, resulting in an improved end-task performance.

While the related literature was scarce at the time of our experiments, several concurrent or later studies have addressed the linguistic capabilities of the multilingual BERT model, providing a great opportunity to independently validate many of our reported results. In particular, de Vries et al. (2020) compare mBERT to a monolingual Dutch BERTje model in probing experiments on POS tagging, dependency parsing and NER. They observe the pipeline-like nature in both models, noting that the monolingual probe consistently relies on later layers compared to the multilingual one – similar to our observation for English and German probes (made, however, on a single multilingual model). They find that the monolingual probes make more use of the earlier layers of the model, and hypothesize that they rely more on lexical information compared to the multilingual probes. In a cross-dataset (but not cross-formalism) comparison, they show tentative evidence that Universal dependency and POS layer utilization does not change much across the datasets.

The work by Chi et al. (2020) focuses on cross-lingual syntactic representation in mBERT and provide evidence that mBERT representations encode syntactic tree distances and learn syntactic representations similar to the Universal Dependencies formalism; they observe that the best representations across all languages can be extracted from the layers 7 and 8 of the mBERT-base, again pointing at the alignment between the representations across languages. Vulić et al. (2020) conduct an extensive lexical-semantic probing study using BERT and mBERT and find that the monolingual lexical representations aggregated from the earlier layers of the model expectedly outperform the multilingual ones.

3.4 Formalisms

We conclude our overview with a discussion of formalisms in NLP and introduce role semantics – a family of linguistic formalisms constituting the linguistic target of our probing study.

3.4.1 Cross-formalism analysis in NLP

Most studies in natural language processing perform a comparison along one or more of the following axes: model, task, dataset, language and formalism. Cross-model studies aim to compare different NLP models and architectures, standard evaluation sets and shared tasks being a prototypical case. Cross-task comparisons compare the performance or behaviour of the same model on a range of tasks. Cross-dataset studies analyze the performance of NLP models on different domains and often deliver insights on out-of-domain robustness of NLP approaches; another wide subclass of studies that fall into this category is dedicated to the robustness of NLP models in low-resource settings. Cross-lingual comparisons investigate how models translate across different languages and typologies. Finally, cross-formalism studies compare across labeling schemes. Cross-model, -domain and -language comparisons are abundant in natural language processing in general and probing in particular: for example, Tenney et al. (2019a) is a cross-task comparison among a range of linguistic probes, the experiments reported in (Devlin et al., 2019) involve a cross-model (BERT vs ELMo) and cross-task (POS and NER) comparison, and the CoNLL-2009 (Hajič et al., 2009) and CoNLL-2018 shared tasks (Zeman et al., 2018) compare dependency parsers in terms of their cross-lingual performance.

Cross-formalism studies are rare in NLP. However, formalism is an important dimension that should be accounted for. As a striking example, comparing dependency parsing and semantic role labeling performance among languages using the CoNLL-2009 shared task data might create an impression that some languages are intrinsically harder to parse than the others and attribute the differences in performance to morphology, word order, etc. However, the CoNLL-2009 datasets differ not only in language, but also in size, domain and formalism. A recent study by Søgaard (2020) shows that once the dependency tasks are cast to the unified UD formalism, the isomorphic graph overlap between training and test sets becomes the second-most important factor predictive of the parsing performance after the dataset size. This kind of insight is only possible when the formalisms across different languages are unified.

The majority of linguistic tasks are only instantiated in one or few linguistic formalisms. One reason behind this is the dataset availability: to be visible for NLP research, a linguistic theory or labeling scheme has to be instantiated in a ready-to-use resource. A prominent example of linguistic pluralism in the NLP task formulation is dependency parsing, which can be represented using Stanford Dependencies, Universal Dependencies, Surface Universal Dependencies (Gerdes et al., 2018), and a plethora of native, language-specific labeling schemes originating from the corresponding linguistic traditions. Other examples of multi-formalism NLP tasks include part-of-speech tagging, morphological analysis, word sense disambiguation and semantic role labeling.

The differences between formalisms can be substantial. Thereby, a probing study showing that a pre-trained BERT model encodes, for example, syntax, in fact demonstrates that the model encodes a particular syntactic formalism (e.g. Universal Dependencies) with a certain degree of success. Does this observation hold for another formalism (e.g. Stanford Dependencies)? Which formalism can the model match better and why? Does the choice of formalism affect the structural probing results? If no, what does it tell us about the formalism and the probing model? If yes, what are the differences and can we explain them based on the formalism definition? Although we are not aware of any large-scale systematic studies dedicated to the effect of formalism on probing results, the evidence of such effects is scattered across the related work: for example, the aforementioned results in Tenney et al. (2019a) show a difference in the layer utilization between constituent- and dependency-based syntactic probes and semantic role and proto-role probes. It is not clear whether this effect is due to the differences in the underlying datasets and task architecture or the formalism per se. In a concurrent recent work, Kulmizev et al. (2020) use a syntactic structural probe by Hewitt and Manning (2019) to compare Universal and Surface-Syntactic Universal Dependencies (SUD) and find that both ELMo and BERT pre-trained models can better fit the Universal Dependency trees. This work shares a lot of our motivation and supports the claim about the importance of cross-formalism studies.

3.4.2 Role Semantics

An ideal setting for a cross-formalism study would encompass several substantially different linguistic formalisms and would isolate the effect of the formalism from the potential dataset and language-related confounds. Role semantics provides a rare opportunity for such a study: decades of research in theoretical linguistics have produced several substantially different role labeling formalisms, and dedicated projects for English and German have created multi-formalism annotations of the same underlying corpora. For further discussion, consider the following synthetic example:

a. *[John]$_{Ag}$ gave [Mary]$_{Rc}$ a [book]$_{Th}$.*

b. *[Mary]$_{Rc}$ was given a [book]$_{Th}$ by [John]$_{Ag}$.*

Despite surface-level differences, the sentences express the same meaning, suggesting an underlying semantic representation in which these sentences are equivalent. One such representation is offered by role semantics – a shallow predicate-semantic formalism closely related to syntax. In terms of role semantics, *"Mary"*, *"book"* and *"John"* are *semantic arguments* of the *predicate "give"*, and are assigned *roles* from a pre-defined inventory, for example, `Agent`, `Recipient` and `Theme`.

Semantic roles and their properties have received extensive attention in linguistics (Fillmore, 1968; Levin and Rappaport Hovav, 2005; Dowty, 1991) and are considered a universal feature of human language. The size and organization of the role and predicate inventory are subject to debate, giving rise to a variety of role-semantic formalisms.

PropBank assumes a predicate-independent labeling scheme where predicates are distinguished by their sense (`get.01`), and semantic arguments are labeled with generic numbered core (`Arg0-5`[4]) and modifier (e.g. `AM-TMP`) roles. Core roles are not tied to specific definitions, but the effort has been made to keep the role assignments consistent for similar verbs; `Arg0` and `Arg1` correspond to the Proto-Agent and Proto-Patient roles as per Dowty (1991). The semantic interpretation of core roles depends on the predicate sense.

VerbNet follows a different categorization scheme. Motivated by the regularities in verb behavior, Levin (1993) has introduced the grouping of verbs into intersective classes (ILC). This methodology has been adopted by VerbNet: for example, the VerbNet class `get-13.5.1` would include verbs *earn*, *fetch*, *gain* etc. A verb in VerbNet can belong to several classes corresponding to different senses; each class is associated with a set of roles and licensed syntactic transformations. Unlike PropBank, VerbNet uses a set of approx. 30 thematic roles that have universal definitions and are shared among predicates, e.g. `Agent`, `Beneficiary`, `Instrument`.

FrameNet takes a meaning-driven stance on the role encoding by modeling it in terms of frame semantics: predicates are grouped into frames (e.g. `Commerce_buy`), which specify role-like slots to be filled. FrameNet offers fine-grained frame distinctions, and roles in FrameNet are frame-specific, e.g. `Buyer`, `Seller` and `Money`. The resource accompanies each frame with a description of the situation and its core and peripheral participants.

[4] An alternative notation spells the role labels as `A0-5`.

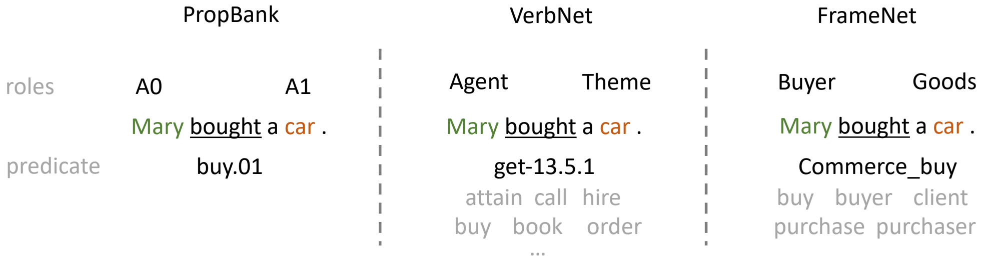


Figure 3.3: Comparison of the major semantic role formalisms in terms of predicate and role groupings. For the sake of presentation, we omit rich additional information, e.g. predicate descriptions, construction sets, selectional restrictions, additional roles etc. provided by each resource.

Figure 3.3 illustrates the differences between the three formalisms. Related research in cross-formalism comparisons for the semantic role labeling tasks demonstrates that each of them offers certain advantages and disadvantages (Giuglea and Moschitti, 2006; Mújdricza-Maydt et al., 2016). While being close to syntax and thereby easier to predict, PropBank doesn't contribute much semantics to the representation. On the opposite side of the spectrum, FrameNet offers rich predicate-semantic representations for verbs and nouns, but suffers from high granularity and coverage gaps (Hartmann et al., 2017c). VerbNet takes a middle ground by following grammatical criteria while still encoding coarse-grained semantics, but only focuses on verbs and core (not modifier) roles (Merlo and van der Plas, 2009).

3.5 Setup

3.5.1 Probe architecture

For our experiments, we take the edge probing setup by Tenney et al. (2019b) as a starting point. Edge probing aims to predict a label given a pair of contextualized span or word encodings. More formally, we encode a WordPiece (WP)-tokenized sentence $[wp_1, wp_2, ...wp_k]$ with a frozen pre-trained model, producing contextual embeddings $[e_1, e_2, ...e_k]$, each of which is a layered representation over $L = \{l_0, l_1, ...l_m\}$ layers, with encoding at layer l_n for the wordpiece wp_i further denoted as e_i^n. A trainable scalar mix is applied to the layered representation to produce the final encoding given the per-layer mixing weights $\{a^0, a^1..a^m\}$ and a scaling parameter γ:

$$\overline{e}_i = \gamma \sum_{l=0}^{m} softmax(a^l) e_i^l$$

Given the source src and target tgt wordpieces encoded as $\overline{e}_{src}$ and $\overline{e}_{tgt}$, our goal is to predict the label y. Figure 3.4 illustrates the architecture of the model.

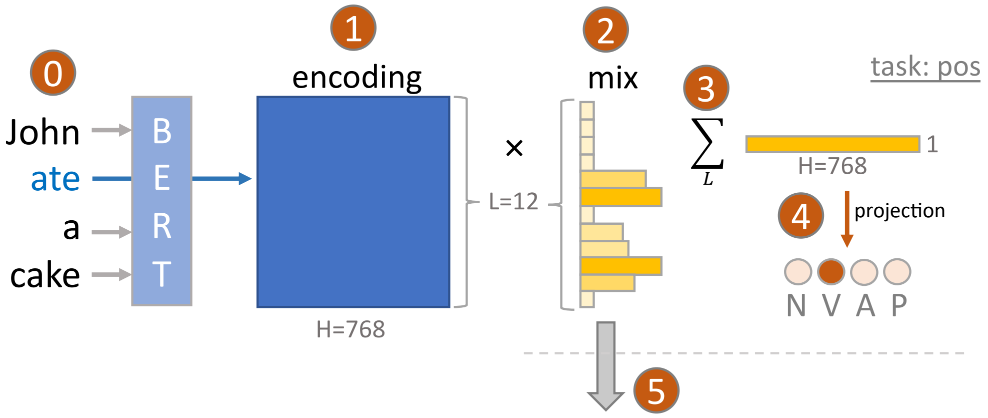


Figure 3.4: Probe architecture with `bert-base` dimensions for illustration. Input sentence (0) is encoded with a frozen BERT model, producing a 12-layer encoding for each input wordpiece (1); each layer in this encoding is multiplied by it's scalar mixing weight (2) and the weighted sum of layers (3) is fed into a linear projection layer (4). Components in blue are frozen, components in orange are updated during training. After training, the scalar mixing weights associated with the task (2) are extracted for further analysis (5) and comparison to other tasks' layer utilizations.

Due to its task-agnostic architecture, edge probing can be applied to a wide variety of unary (by omitting *tgt*) and binary labeling tasks in a unified manner, facilitating the cross-task comparison. The original setup has several limitations that we address in our implementation.

Regression tasks. The original edge probing setup only considers classification tasks. Many language phenomena – including positional information, distances and syntactic depth, are naturally modeled as regression. We extend the architecture by Tenney et al. (2019b) and support both classification and regression: the former is achieved via softmax, the latter works via direct linear regression to the target value.

Flat model. To decrease the models' own expressive power (Hewitt and Liang, 2019), we keep the number of parameters in our probing model as low as possible. While Tenney et al. (2019b) utilize pooled self-attentional span representations and a projection layer to enable cross-model comparison, we directly feed the wordpiece encoding into the classifier, using the first wordpiece of a word. To further increase the selectivity of the model, we directly project the source and the target wordpiece representations into the label space, opposed to the two-layer multi-layer perceptron classifier used in the original setup.

Separate scalar mixes. To enable fine-grained analysis of probing results, we train and analyze separate scalar mixes for source and target wordpieces, motivated by the fact that the classifier might utilize different aspects of their representation for prediction.[5] Indeed, we find that the mixing weights learned

[5] Tenney et al. (2019b, Appendix C) also use separate mixes in the background, but do not

for source and target wordpieces might show substantial – and linguistically meaningful – variation. Combined with a regression-based objective, separating the scalar mixes allows us to analyze layer utilization patterns for the semantic proto-role features in Chapter 4.

Sentence-level probes. Utilizing the BERT-specific sentence representation *[CLS]* allows us to incorporate the sentence-level natural language inference (NLI) probe into our kit.

Anchor tasks. We employ two analytical tools from the original layer probing setup. Mixing weight plotting compares layer utilization among tasks by visually aligning the respective learned weight distributions transformed via a softmax function. Layer center-of-gravity is used as a summary statistic for a task's layer utilization. While the distribution of mixing weights along the layers allows us to estimate the order in which the information is processed during encoding, it doesn't allow us to directly assess the *similarity* between the layer utilization of the probing tasks.

Tenney et al. (2019a) have demonstrated that the order in which linguistic information is stored in BERT mirrors the traditional NLP pipeline. A prominent property of the NLP pipelines is their use of low-level features to predict downstream phenomena. In the context of layer probing, probing tasks can be seen as end-to-end feature extractors. Following this intuition, we define two groups of probing tasks: *target tasks* – the main tasks under investigation, and *anchor tasks* – a set of related tasks that serve as a basis for qualitative comparison between the target tasks. The softmax transformation of the scalar mixing weights allows us to treat them as probability distributions: the higher the mixing weight of a layer, the more likely the probe is to utilize the information from this layer during prediction. We use Kullback-Leibler divergence to compare target tasks (e.g. role labeling in different formalisms) in terms of their similarity to lower-level anchor tasks (e.g. dependency relation and lemma). Note that the notion of an anchor task is contextual: the same task can serve as a target and as an anchor, depending on the focus of a study.

3.5.2 Source data

For German, we use the SR3de corpus (Mújdricza-Maydt et al., 2016) that contains parallel PropBank, FrameNet and VerbNet annotations for verbal predicates. For English, SemLink (Bonial et al., 2013; Stowe et al., 2021) provides mappings from the original PropBank corpus annotations to the corresponding FrameNet and VerbNet senses and semantic roles. We use these mappings to enrich the CoNLL-2009 (Hajič et al., 2009) dependency role labeling data – also based on the original PropBank – with roles in all three formalisms via a semi-automatic token alignment procedure. We only accept verbal predications from CoNLL-2009 where all original semantic roles could be unambiguously aligned to their VerbNet and FrameNet counterparts via

investigate the differences between the learned layer utilizations.

	tok	sent	pred	arg
CoNLL+SL	312.2K	11.3K	13.3K	23.9K
SR3de	62.6K	2.8K	2.9K	5.5K

Table 3.1: Statistics for CoNLL+SemLink (English) and SR3de (German), only core roles.

SemLink. This results in a substantial data size reduction, since VerbNet only specifies role-semantic information for verbs and core roles; in addition, not all predications could be fully aligned.

The resulting corpus is smaller than the original one, but still is an order of magnitude larger than SR3de (Table 3.1). Both corpora are richly annotated with linguistic phenomena on word level, including part of speech, lemma and syntactic dependencies. The natural language inference (NLI) probe is sourced from the corresponding development split of the XNLI (Conneau et al., 2018b) dataset.

	type	en	de
`*token.ix`	unary	208.9K	46.9K
`ttype` [v]	unary	177.2K	34.0K
`lex.unit` [v]	unary	187.6K	35.7K
`pos`	unary	312.2K	62.6K
`deprel`	binary	300.9K	59.8K
`role`	binary	23.9K	5.5K
`xnli`	unary	2.5K	2.5K

Table 3.2: Probing task statistics. Tasks marked with [v] use a most frequent label vocabulary. Here and further, tasks marked with * are regression tasks.

3.5.3 Probing kit

Our probing kit spans a wide range of probing tasks, from primitive surface-level tasks mostly utilized as anchors later to high-level semantic tasks that aim to provide a representational upper bound to predicate semantics. We follow the training, test and development splits from the original SR3de and CoNLL-2009 data. The XNLI task is sourced from the development set and only used for scalar mix analysis. To reduce the number of labels in some of the probing tasks, we collect frequency statistics over the corresponding training sets and

language	en	de
PropBank	5	10
VerbNet	23	29
FrameNet	189	300

Table 3.3: # of role probe labels by formalism.

task	input	label
`*token.ix`	I [saw] a cat.	→ 2
`ttype`	I [saw] a cat.	→ saw
`lex.unit`	I [saw] a cat.	→ see.V
`pos`	I [saw] a cat.	→ VBD
`deprel`	$[\text{I}]_{\text{tgt}}$ $[\text{saw}]_{\text{src}}$ a cat.	→ SBJ
`role.vn`	$[\text{I}]_{\text{tgt}}$ $[\text{saw}]_{\text{src}}$ a cat.	→ Experiencer

Table 3.4: Word-level probing task examples for English.

only consider up to 250 most frequent labels. Below we define the tasks in order of their complexity, Table 3.2 provides the probing task statistics, Table 3.3 compares the categorical role labeling formalisms in terms of granularity, and Table 3.4 provides examples. We evaluate the classification performance using Accuracy, while regression tasks are scored via R^2.

Token position (`token.ix`) predicts the linear position of a word, cast as a regression task over the first 20 words in the sentence. Again, the task is non-trivial since it requires the words to be assembled from the wordpieces.

Token type (`ttype`) predicts the type of a word. This requires contextual processing since a word might consist of several wordpieces;

Lexical unit (`lex.unit`) predicts the lemma and POS of the given word – a common input representation for the entries in lexical resources. We extract coarse POS tags by using the first character of the language-specific POS tag.

Part of speech (`pos`) predicts the language-specific part-of-speech tag for the given token.

Dependency relation (`deprel`) predicts the dependency relation between the parent `src` and dependent `tgt` tokens;

Semantic role (`role.[frm]`) predicts the semantic role given a predicate `src` and an argument `tgt` token in one of the three role labeling formalisms: PropBank `pb`, VerbNet `vn` and FrameNet `fn`. Note that we only probe for the role label, and the model has no access to the verb sense information from the data.

XNLI is a sentence-level natural language inference (NLI) task directly sourced from the corresponding dataset. Given two sentences, e.g. *"John is reading a book"* and *"John is sleeping"* the goal is to determine whether an entailment or a contradiction relation holds between them (in this case, "contradiction"). We use NLI to investigate the layer utilization of mBERT for high-level semantic tasks. We extract the sentence pair representation via the *[CLS]* token and treat it as a unary probing task.

task	en	de
`*token.ix`	0.95 (0.93)	0.92 (0.87)
`ttype`	1.00 (0.92)	1.00 (0.48)
`lex.unit`	1.00 (0.75)	1.00 (0.33)
`pos`	0.97 (0.40)	0.97 (0.26)
`deprel`	0.95 (0.42)	0.95 (0.41)
`role.fn`	0.92 (0.18)	0.59 (0.10)
`role.pb`	0.96 (0.67)	0.71 (0.49)
`role.vn`	0.94 (0.47)	0.73 (0.30)

Table 3.5: Best dev score for word-level tasks over 20 epochs, *Acc* for classification, R^2 for regression; Baseline in parentheses.

3.5.4 Implementation

Our probing framework is implemented using AllenNLP.[6] We train the probes for 20 epochs using the Adam optimizer with default parameters and a batch size of 32. Due to the frozen encoder and flat model architecture, the total runtime of the main experiments is under 8 hours on a single Tesla V100 GPU. In addition to pre-trained mBERT, we report baseline performance using a frozen untrained mBERT model obtained by randomizing the encoder weights post-initialization using the method from Jawahar et al. (2019).

3.6 Main results

3.6.1 General Trends

While absolute performance is secondary to our analysis, we report the probing task scores on the respective development sets in Table 3.5. We observe that grammatical tasks score high, while core role labeling lags behind – in line with the findings of Tenney et al. (2019a).[7] We observe lower scores for German role labeling which we attribute to the lack of training data. Surprisingly, as we show below, this doesn't prevent the edge probe from learning to locate relevant role-semantic information in mBERT's layers.

The untrained mBERT baseline expectedly underperforms. We note good baseline results on surface-level tasks for English, which we attribute to memorizing token identity and position: although the weights are set randomly, the frozen encoder still associates each wordpiece input with a fixed random vector. We have confirmed this assumption by scalar mix analysis of the untrained mBERT baseline: in our experiments, the baseline probes for both English and German attended almost exclusively to the first few layers of the encoder, independent of the task. We attribute the lower baseline scores for

[6] Code available: `https://github.com/UKPLab/emnlp2020-formalism-probing`

[7] Our results are not directly comparable due to the differences in datasets and formalisms.

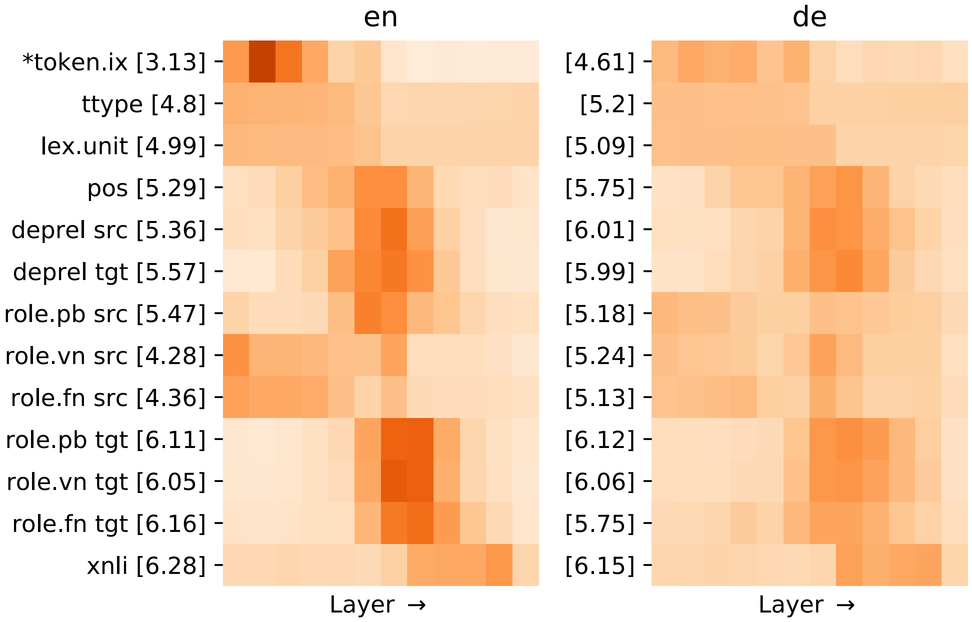


Figure 3.5: Layer probing results by task, the center-of-gravity statistic in square brackets.

German to the differences in dataset size, which play an increasing role as the randomized baseline encoder no longer supplies the relevant information learned during pre-training. For brevity, here and further we do not examine baseline mixing weights and only report the scores.

The big picture of layer utilization in Figure 3.5 mirrors the findings of Tenney et al. (2019a) about the sequential processing order in BERT. We observe that the layer utilization among tasks generally aligns for English and German, although we note that in terms of center-of-gravity mBERT tends to utilize deeper layers for German probes.[8] Basic word-level tasks are indeed processed early by the model, and XNLI probes focus on deeper levels, suggesting that the representation of higher-level semantic phenomena follows the encoding of syntax and predicate semantics. A possible confounding factor in case of XNLI is the sentence-pair nature of the task; we revisit this observation in Section 3.7.3.

3.6.2 The Effect of Formalism

Using separate scalar mixes for source and target tokens (e.g. *"saw"* and *"cat"* in Table 3.4) allows us to explore the cross-formalism encoding of role semantics by mBERT in detail. For both English and German role labeling, the probe's layer utilization drastically differs for predicate and argument tokens.

[8] A recent study by de Vries et al. (2020) reports a similar trend when comparing a language-specific and a multilingual BERT model for Dutch.

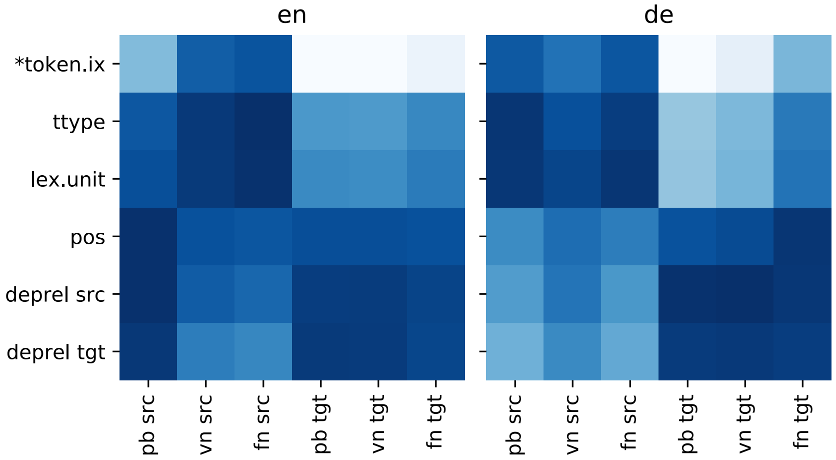


Figure 3.6: Anchor task analysis of SRL formalisms.

While the argument representation `role*tgt` mostly focuses on the same layers as the dependency parsing probe, the layer utilization of the predicates `role*src` is affected by the chosen formalism. In English, PropBank predicate token mixing weights emphasize the same layers as dependency parsing – in line with the previously published results. However, the probes for VerbNet and FrameNet predicates (`role.vn src` and `role.fn src`) utilize the layers associated with `ttype` and `lex.unit` that contain lexical information. Concurrent work by Vulić et al. (2020) validates our intuition about the localization of lexical information. Coupled with the fact that both VerbNet and FrameNet assign semantic roles based on lexical-semantic predicate groupings (frames in FrameNet and verb classes in VerbNet), this suggests that the lower layers of mBERT implicitly encode predicate sense information; moreover, sense encoding for VerbNet utilizes deeper layers of the model associated with syntax[9], in line with VerbNet's predicate classification strategy. This finding confirms that the formalism can indeed have linguistically meaningful effects on probing results.

3.6.3 Anchor Tasks and the Pipeline

We now use the scalar mixes of the role labeling probes as target tasks, and lower-level probes as anchor tasks to qualitatively explore the differences between how our role probes learn to represent predicates and semantic arguments[10] (Figure 3.6). The results reveal a distinctive pattern that confirms our previous observations: while VerbNet and FrameNet predicate layer utilization `src` is similar to the scalar mixes learned for `ttype` and `lex.unit`, the learned argument representations `tgt` and the PropBank predicate `pb src` attend to

[9] Layers 7-8 of mBERT base, similar to the new reports by Chi et al. (2020) and Kulmizev et al. (2020)

[10] Darker color corresponds to higher similarity.

the layers associated with dependency and POS probes. Aside from the PropBank predicate encoding which we address below, the pattern reproduces for English and German. This aligns with the traditional separation of the semantic role labeling task into predicate disambiguation followed by semantic argument identification and labeling, along with the feature sets employed for these tasks (Björkelund et al., 2009). Note that the observation about the pipeline-like task processing within the BERT encoders thereby holds, albeit on a sub-task level.

3.6.4 Formalism Implementation

Both layer and anchor task analyses reveal a prominent discrepancy between English and German role probing results: while the PropBank predicate layer utilization for English mostly relies on syntactic information, German PropBank predicates behave similarly to VerbNet and FrameNet. The lack of systematic cross-lingual differences between layer utilization for other probing tasks[11] allows us to rule out the effect of purely typological features such as word order and case marking as a likely cause.

The difference in the number of role labels for English and German PropBank, however, points at possible qualitative differences in the labeling schemes (Table 3.3). The data for English stems from the token-level alignment in SemLink that maps the original PropBank roles to VerbNet and FrameNet. Role annotations for German have a different lineage: they originate from the FrameNet-annotated SALSA corpus (Burchardt et al., 2006) semi-automatically converted to PropBank style for the CoNLL-2009 shared task (Hajič et al., 2009), and enriched with VerbNet labels in SR3de (Mújdricza-Maydt et al., 2016). While English PropBank *labels* only weakly depend on the predicate identity (although their predicate-specific *interpretation* might differ), the German dataset conversion procedure described in (Hajič et al., 2009) suggests that German PropBank, following the same numbered labeling scheme, keeps this scheme *consistent within the frame* (Figure 3.7). We assume that this latent grouping incentivizes the probe to leverage the lexical-semantic information for predicates, and reflects in our probing results. The ability of the probe to detect subtle differences between formalism implementations would constitute a new use case for probing, and a promising direction for future studies.

3.7 Supplementary experiments

Although determining the impact of a formalism is our main focus, we now make several general observations about the layer probing behavior in the context of cross-formalism studies. The layer importance weights for German are more uniform than the weights learned on English data. The German dataset

[11] Apart from the general tendency to use deeper layers in German reported above in Section 3.6.1.

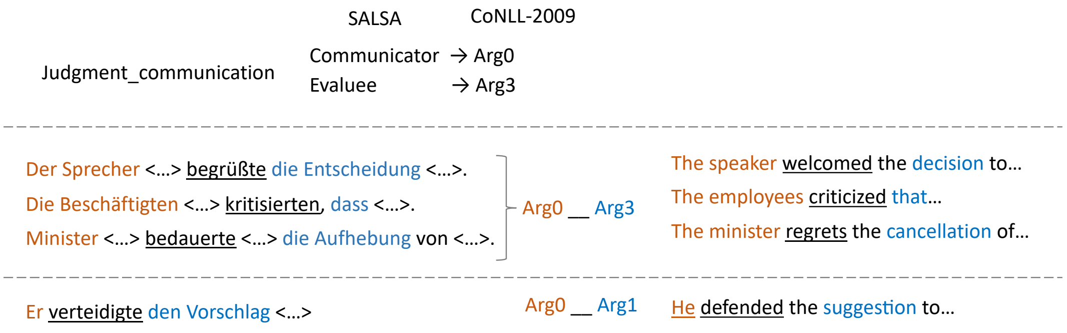

SALSA
CoNLL-2009
Judgment_communication
Communicator → Arg0
Evaluee → Arg3
Der Sprecher <...> begrüßte die Entscheidung <...>.
Die Beschäftigten <...> kritisierten, dass <...>.
Minister <...> bedauerte <...> die Aufhebung von <...>.
Arg0 __ Arg3
The speaker welcomed the decision to...
The employees criticized that...
The minister regrets the cancellation of...
Er verteidigte den Vorschlag <...>
Arg0 __ Arg1
He defended the suggestion to...

Figure 3.7: PropBank roles in German CoNLL-2009 and SR3de data (simplified excerpts with English translations). Because the three predicates originally belonged to the same frame, their direct objects are consistently assigned the same `Arg3` PropBank role according to the frame-level mapping; while the object of a predicate from another frame gets a different PropBank role assignment – `Arg1`.

is also an order of magnitude smaller; can this difference be attributed to the dataset size? Section 3.7.1 sheds light on the behavior of layer probes in low-resource scenarios. The data for English and German comes from different datasets and domains, and yet the general layer probing results align. Does this mean that the dataset per se is of secondary importance? We investigate this in an additional experiment on the data from four Universal Dependency treebanks in Section 3.7.2. Finally, although most layer probing studies focus on word-level tasks, our results for the sentence-level XNLI suggest that high-level semantic processing happens in the later layers of the model. However, can it be that the sentence-level encoding or the sentence pair task formulation are responsible for this effect? We conclude our investigation with an experiment on additional sentence-level tasks in Section 3.7.3.

3.7.1 The impact of data size

Since parallel multi-formalism annotations are rare, tasks in cross-formalism probing studies would often need to be sourced from different corpora; and the corpus difference is almost guaranteed in a multilingual setup. One of the most basic properties of a corpus is its size: a larger corpus provides more training signal and enables better generalization. In our study, we have observed that the German layer utilization is more uniformly distributed across the layers (corresponding to the absense of bright "peaks" in Figure 3.5) compared to English. We hypothesize that the size difference between English and German data is partly responsible for this effect. To verify this hypothesis, we conduct an additional experiment by reducing the English training data size: a probe is now trained on a subset of $K = [100, 500, 1000, 3000, 5000]$ training sentences, and evaluated according to the protocol used in the main study. The $K = 3000$ setting roughly corresponds to the German SR3de training set conditions.

Figure 3.8 compares the performance of edge probes for `pos`, `deprel` and role

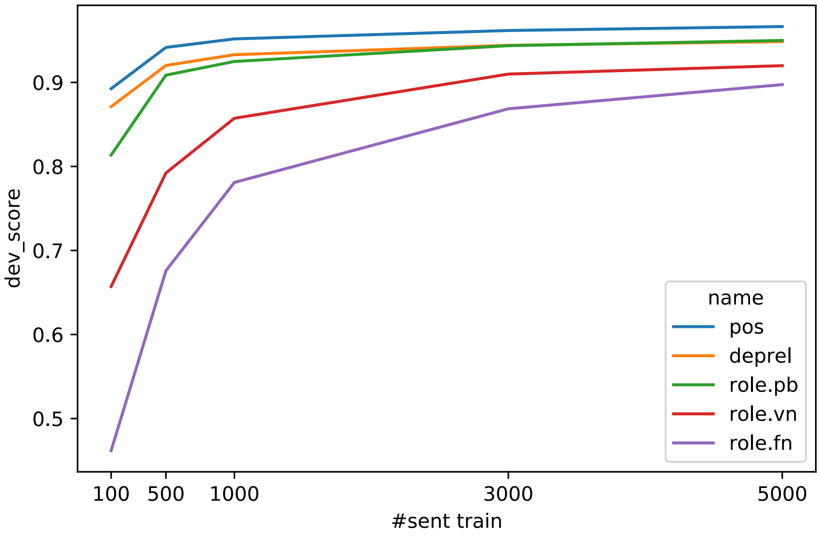


Figure 3.8: English dev *Acc* depending on training data size, by task.

labeling tasks. We can observe the differences in the amount of data needed to achieve high performance depending on the task: while `pos`, `deprel` and PropBank role labeling probe plateau after having been exposed to 1000-2000 sentences, VerbNet and FrameNet probes gradually improve their performance as more data gets available. Figure 3.9, in turn, displays the differences between measured layer utilization depending on the amount of training data available. Confirming our hypothesis, less training data does not only lead to lower performance, but also results in more uniformly distributed scalar mixing weights; however, the coarse layer utilization pattern already emerges after 500-1000 sentences depending on the task and gradually becomes more peaked as more data is added, converging to the patterns similar to the full-data experiment seen in Figure 3.5.

This preliminary finding has three important implications. First, it allows us to attribute part of the difference in general German and English probing results to the dataset size difference. Second, it suggests that at least from the layer probing perspective, a relatively small data sample can be sufficient for performing cross-formalism probing – which is promising, given that cross-formalism annotation is a time-consuming task that requires linguistic expertise. Finally, this allows an alternative interpretation of uniform layer utilization for a task: while one might see this as a signal of the linguistic information being distributed throughout the model (Tenney et al., 2019a), a possible alternative is that the probing model simply has not observed enough data to localize the parts of the pre-trained encoder most suited to solve the task.

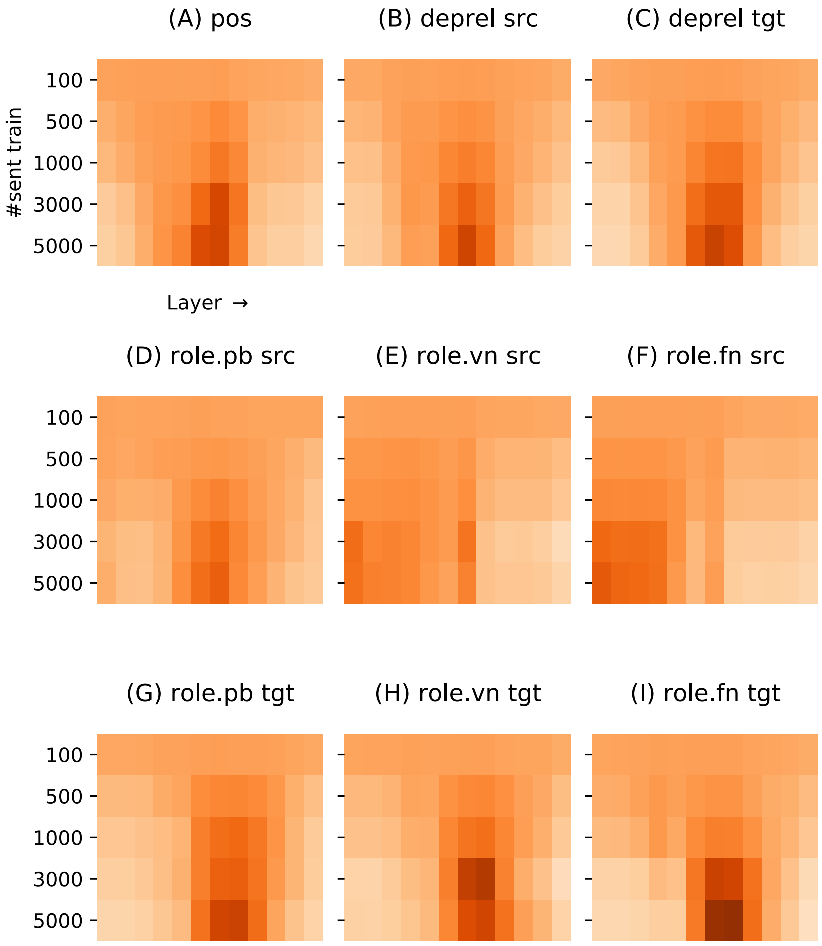


Figure 3.9: Layer utilization depending on training data size.

3.7.2 The impact of dataset

One of the motivations behind our choice of the main probing task – role labeling – is the availability of parallel cross-formalism annotations. This arrangement, however, is rare, and we observe a satisfactory level of alignment between English and German results despite the dataset difference. A similar observation is made in concurrent work by de Vries et al. (2020), who report that the part-of-speech and dependency probing results on Dutch UDLassy and UDAlpino datasets align and conclude that the probes are "sensitive to the task and the input embeddings, but not overly sensitive to the specific data that the probes are trained on". This is intriguing, as the alignment between the probing results for different datasets would allow cross-formalism studies based on different corpora, substantially relaxing the requirements for such studies.

To investigate this, we depart from our original data and probing kit and conduct additional experiments on four Universal Dependencies treebanks for

English:

- English Web Treebank (EWT, Silveira et al. (2014)) is a collection of texts from various web media types incl. blogs, e-mails, question-answering platforms, product reviews and newsgroups.
- The Georgetown University Multilayer corpus (GUM, Zeldes (2017)) is a collection of freely available texts from the web covering diverse text types from the academic, fiction, non-fiction, spoken and newswire domains.
- Parallel TUT from the University of Turin (ParTUT. (Bosco et al., 2012)) is a collection of parallel Italian, English and French sentences sourced from Wikipedia, legal texts and public talks.
- LinES (Ahrenberg, 2015) is a conversion of the parallel English-Swedish treebank covering several literary works and Europarl data.

All four corpora are annotated with the standard Universal Dependencies layers; the source data was extracted from the public distribution of Universal Dependencies v. 2.7. We generate a probing kit based on this data and the tasks used in the main study; note that unlike our main probing kit, here we probe for *universal* part-of-speech tags and dependencies, as opposed to language-specific schemata used in CoNLL-2009SL and SR3de. This way the formalism and language are fixed, and the only difference between the experimental runs is the underlying dataset. We include the token type prediction probe as a baseline sanity check.

Table 3.6 provides dataset statistics and summarizes the probing task performance. As we can observe, the datasets differ in size; however, the probing performance does not seem to be greatly affected by this, and all the probes show high development accuracy. We note, however, that this might be due to the task selection, as both dependency relation and part-of-speech tagging performed well on limited data in our previous experiment.

Figure 3.10 compares the scalar mixes learned by the probes across tasks and datasets. We can observe that the general layer utilization patterns and the pipeline-like order of the tasks indeed transfer well across datasets. The concentration of the scalar mixing weight on a few layers is in line with our previous observations: as the training set size becomes smaller (EWT → GUM → LINES → PARTUT) the layer weights are more uniformly distributed.

This preliminary finding, mirrored by the observations in a related study from de Vries et al. (2020), is promising, since the lack of dataset dependency would make cross-formalism studies easier to set up. However, we point out that both our and the concurrent results were obtained for the layer probing setup, and it remains unclear to which extent this observation generalizes to other probing methodologies, models, datasets and tasks. We leave the exploration of this question to future work.

dataset	#s train	deprel	ttype	upos
ewt	12k	0.93	1.00	0.95
gum	4k	0.93	0.99	0.96
lines	3k	0.92	1.00	0.97
partut	1.8k	0.92	1.00	0.95

Table 3.6: UD Dataset size and probing performance (dev *Acc*)

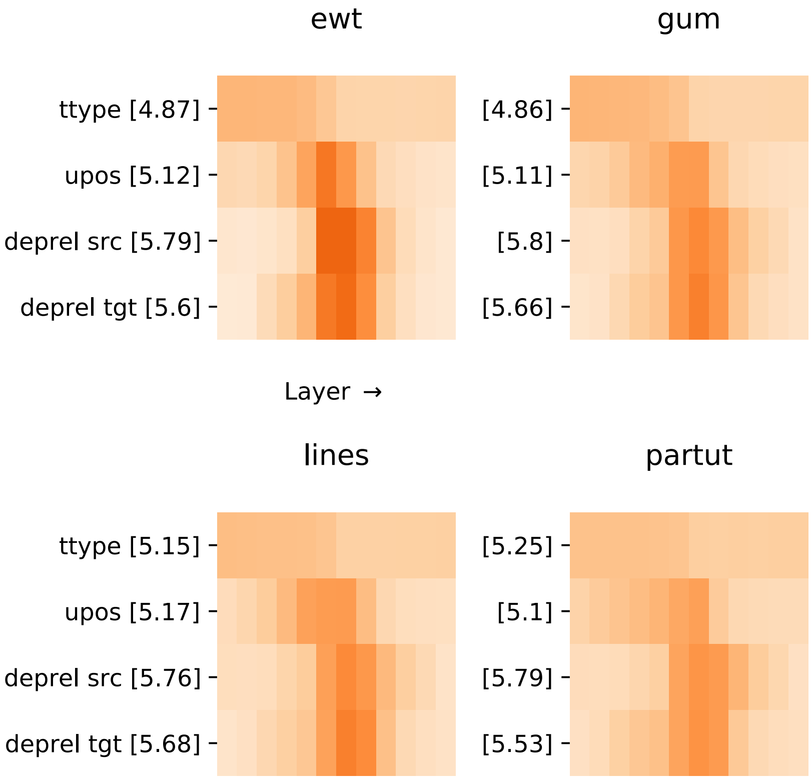


Figure 3.10: Layer utilization across UD datasets.

3.7.3 Layer probing on sentence level

Unlike the work by Tenney et al. (2019a), our edge probing kit incorporates the sentence-level XNLI task, which is formulated as a unary probe using the special *[CLS]* token. Our main probing results in Figure 3.5 suggest that the XNLI probe learns to prioritise later layers of the pre-trained mBERT model, suggesting access to high-level semantic abstractions. However, XNLI is the only sentence-level probe in our kit. Can this effect be due to the sentence-level task formulation or the use of the special classification token as the representation of a sentence? To investigate this, we further extend our kit with two basic sentence-level tasks:

Sentence Length (`senlen`) is a regression task of predicting the length of the sentence in terms of word pieces. One could expect that this basic operation would require access to the lower layers of the model since the sequence length can be retrieved as an index of the last token in the sequence. With this task, we investigate the effect of sentence-level encoding on the layer probing results.

Pairwise Length Comparison (`pairlen`) is a binary classification task that compares the length of two input sentences (encoded jointly using the *[SEP]* token; it labels an instance as positive if the first sentence in the pair is longer than the second one. With this task, we investigate the pairwise sentence encoding and its effects on layer probing.

Both new tasks are sourced from the same XNLI development split, which is also used for the evaluation. Sentences in XNLI sentence pairs are *ordered* and differ in status and origin. The first sentence in the English XNLI pair is a premise extracted from a pre-existing text. The second sentence is a hypothesis for the given premise collected via crowdsourcing. The multilingual portion of the XNLI is then manually translated from English. Our analysis of the dataset has revealed that a naive implementation of the `pairlen` probe as described above would suffer from label imbalance, as the strong majority of the first (naturally occurring) sentences are longer than the second (crowd-generatd) sentences.[12] To prevent the probe from exploiting the potential confound of sentence position ("first sentence is always longer"), we randomly swap the first and second sentence in the XNLI pairs for the `pairlen` probe, achieving a balanced distribution and ensuring that the probe would indeed need to learn to compare sentence lengths.

The performance of the probes is shown in Table 3.7, while Figure 3.11 compares the scalar mixing weights for sentence-level tasks in English and German. Not unexpectedly, XNLI turns out to be the most challenging task out of the three.[13] Turning to the scalar mixes, while the XNLI probe indeed uses the

[12] This is true for 84% cases in English and 82% in German XNLI development data.

[13] Note that here we evaluate and test on the same development split, as XNLI offers no training set. The scores therefore only illustrate how well the probe was able to fit the training data, and not how well it would perform on new data.

task	target	en	de
*senlen	single	0.885	0.880
pairlen	pair	0.941	0.943
xnli	pair	0.613	0.659

Table 3.7: Sentence probing performance, R^2 for `senlen`, *Acc* for others.

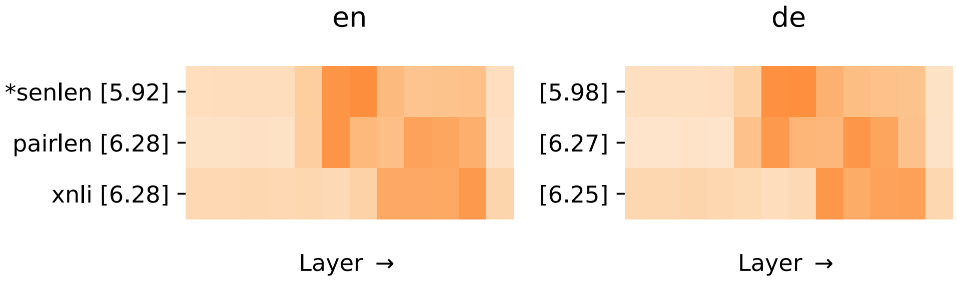


Figure 3.11: Scalar mixing weights for sentence-level probes

later layers of the model than the two basic tasks, the sentence length probes learn to attend to the middle layers of the model. However, the pairwise length comparison probe also attends to the later layers 8-10. This allows us to assume that while the model might indeed attempt to use higher-level semantic information for solving the actual XNLI task, at least some of its layer utilization should be attributed to the pairwise nature of the task. Surprisingly, the sentence length probe does not learn to attend to the same layers as the `token.ix` positional probe; we hypothesize that this is due to the use of the *[CLS]* token and not the actual last wordpiece of the sentence. As sentence embeddings produced at the *[CLS]* have generally been shown to underperform (Reimers and Gurevych, 2019), the exploration of alternative sentence encoding strategies (e.g. averaging over individual wordpiece encodings) could provide deeper insights into the layer utilization for sentence-level tasks. We leave this investigation to the future.

3.8 Outlook

The impact of formalism. We have demonstrated that the choice of linguistic formalism can have substantial, linguistically meaningful effects on role-semantic probing results. We have shown how probing classifiers can be used to detect discrepancies between formalism implementations. Our refined implementation of the edge probing framework coupled with the anchor task methodology enabled new insights into the processing of predicate-semantic information within mBERT. Our findings suggest that the linguistic formalism is an important factor to be accounted for in probing studies. A more subtle result highlights the importance of the probing architecture: while our role probe reflects the differences between PropBank, FrameNet and VerbNet formalisms, the layer utilization of the syntactic dependency tasks seems to align between formalisms and datasets. Kulmizev et al. (2020) show that using a different, structural probe can detect the differences between syntactic formalisms, and analyze the reasons behind those differences. It is our hope that our and concurrent work motivate further research in cross-formalism NLP and probing. We stress the importance of negative results for future work: if two substantially different formalisms result in the same probing measurements, it is important to understand what it tells us about the probe, the task and the formalisms.

Recommendations. Our finding prompts several general recommendations for the follow-up probing studies. First, the formalism and implementation used to prepare the linguistic material underlying a probing study should be always explicitly specified. Second, if possible, results on multiple formalisations of the same task should be reported and validated for several languages. Assembling corpora with parallel cross-formalism annotations would facilitate further research on the effects of a formalism in probing; however, our preliminary cross-dataset results suggest that strictly parallel data might not be a hard pre-requisite for such studies, as discussed below.

Requirements for cross-formalism studies. In a range of supplementary experiments, we have explored the properties of layer probes with respect to the data size, dataset and task inputs. We have shown that the amount of training data can have substantial effect on layer probing results; however, just a few thousand sentences might be enough to get an overall picture of the layer utilization by a given task. This is promising as it would reduce the amount of formalism-specific data the expert linguists would need to annotate; on the other hand, we point out the importance of validating this observation for other tasks and probing methodologies, and highlight the associated dangers of losing long-tail phenomena due to the dataset size reduction. Our further results show that the dataset itself does not have a prominent effect on the layer utilization patterns as measured by our probe; if true in general, this would reduce the requirements for cross-formalism studies even further as it would allow comparing the formalisms across datasets. The effect of dataset and domain on probing is a compelling target for further research, including more tasks, probing methodologies and domains. Finally, in our experiments on sentence-level inputs we have found that even basic, count-based sentence pair tasks tend to use the later layers of the model; further investigation using alternative sentence encoding strategies would shed more light at the layer-wise encoding of sentence-level phenomena in pre-trained models.

Probing methodology. The area of probing and interpretability is rapidly advancing, and recent studies open new methodological possibilities for extending our work. Driven by the criticism of the classical probing classifier setup, Voita and Titov (2020) present an alternative, information-theoretic probe; and Elazar et al. (2021) suggest a way to bridge layer probing results with the model's behavior on end tasks. Finally, Ribeiro et al. (2020) propose a test-driven approach to NLP model evaluation that naturally lends itself to applications in probing. Exploring the effects of a linguistic formalism on probing results using the newly proposed probe architectures, benchmark construction kits and datasets is an exciting avenue for future research.

3.9 Chapter Summary

In this chapter, we have explored the effects of linguistic formalism choice on the probing results for strong contextualized pre-trained encoders. We have reviewed the existing approaches to contextualization and model analysis, and conducted extensive experiments on role-semantic probing for three formalisms – PropBank, VerbNet and FrameNet – and two languages, English and German.

We have extended the previously proposed layer probing architecture to enable fine-grained insights into the layer utilization in BERT models. We have proposed anchor tasks – a new analytical tool for visualizing similarities between layer probes, drawing inspiration from the feature-based NLP models and the pipeline-like nature of information processing in the BERT layers.

Our main result suggests that the formalism can have a substantial effect on probing measurements. We complement it with a range of additional experiments that provide deeper insights into the layer probing behavior in low-data, cross-dataset and sentence-level settings. A range of recommendations and future work directions conclude the chapter.

Chapter 4

Probing for Role Prominence

In Chapter 2, we examined how linguistic conceptualization can affect NLP evaluation by exploring the mismatch between static word embeddings and the evaluation targets in lexical similarity benchmarks and resource-based scenarios. Chapter 3 explored the contribution of linguistics to probing and examined the effects of linguistic formalisms on probing results. In the final chapter of this thesis, we turn to the impact of linguistics on probing task design by exploring an alternative, regression-based formulation of the semantic role labeling task, and applying this formulation to probe pre-trained encoders for sentence-level role prominence.

4.1 Introduction

Boosted by the release of large-scale categorical datasets like PropBank (Palmer et al., 2005a) and FrameNet (Baker et al., 1998), role semantics in NLP has been almost exclusively modelled as a class-based semantic role labeling task (Gildea and Jurafsky, 2000). However, there is an alternative, *prominence*-based view on role semantics, which sees roles not as independent categories, but as members of a *hierarchy* that determines the restrictions on the syntactic realization of the roles. In other words, while most work in NLP treats role semantics as classification over a discrete set of labels, it can be cast as regression over a continuous prominence spectrum; as we show, this makes it possible to evaluate the linguistic capacity of pre-trained contextualized encoders on a predication level, instead of mere local role assignments.

In this Chapter, we probe the pre-trained mBERT model for Dowty's Proto-Role properties, and find that – contrary to the existing evidence – some of those properties can be localized in mBERT layer structure. We proceed by introducing a regression-based approach to role probing, and use it in a novel prominence probing setup based on PropBank data; our evidence suggests that mBERT indeed encodes prominence relations between semantic roles. Thematic hierarchies are a popular linguistic device to account for the prominence relations between roles; however, the composition of the universal thematic hierarchy and its mere feasibility are subject to debate. We conclude the thesis with a non-neural approach and report the first attempt to operationalize thematic hierarchies in NLP based on manually annotated syntactic dependency trees. We show that hierarchies resembling the proposals in literature can be extracted from corpus data in an efficient manner and to some extent transfer cross-lingually, and highlight the challenges associated with computational modeling of thematic hierarchies.

4.1.1 Motivation

Before we proceed, we revisit our discussion of role-semantic frameworks from Chapter 3, this time in the context of global relationships between semantic arguments, and prominence. Note that while theoretical linguistics offers a vast body of research on prominence and argument realization, for the purpose of this work we resort to a shallow, simplified notion of role prominence as a *hierarchy that ranks semantic roles on a linear scale based on their importance and associated syntactic prominence*. The recent summary by Levin (2019) provides a great entry point for a detailed discussion of prominence from the linguistics perspective.

Categorical semantic roles in the modern sense were introduced in the 1960s as a way to account for variation in the syntactic behavior of verbs which cannot be explained by purely syntactic means (Gruber, 1965; Fillmore, 1968). A commonly used motivational example contrasts the use of verbs *hit* and *break*: while both are regular transitive verbs, *hit* does not allow construction (4); and construction (5) is ungrammatical in both cases.

(1) [John]$_\text{X}$ broke/hit the [window]$_\text{Y}$ with a [stone]$_\text{Z}$.

(2) [John]$_\text{X}$ broke/hit the [window]$_\text{Y}$.

(3) A [stone]$_\text{Z}$ broke/hit the [window]$_\text{Y}$.

(4) The [window]$_\text{Y}$ broke/*hit.

(5) The [window]$_\text{Y}$ *broke/*hit with a [stone]$_\text{Z}$.

There exist several principled ways to describe the syntactic behavior of arguments in the lexicon. Available constructions can be defined individually on a *verb sense* basis. This strategy is precise but very inefficient, since verbs show substantial similarities in the way they encode their semantic arguments.

A step towards a more general representation is *verb class* grouping (Levin, 1993): verb senses can be grouped into verb classes with syntactic behavior shared among the members of the class. For example, syntactically *break* behaves like *crash*, *shred* and *split*, while *hit* behaves like *bash* and *whack* in the corresponding verb senses. This significantly reduces the lexicon redundancy and allows treatment of the out-of-vocabulary verbs if the verb class can be determined. A similar level of granularity is used by the major role labeling frameworks, FrameNet (Baker et al., 1998) and, to some extent, PropBank (Palmer et al., 2005a).

Semantic arguments share similarities across verb classes, giving rise to the notion of general semantic roles. While there exists no consensus on the inventory of semantic roles, a subset shared by the most theoretical approaches includes roles such as `Agent` (the active sentient initiator of the event), `Theme` (the most affected participant), `Result` (the outcome of the event), `Instrument` (the instrument used) etc. Semantic roles show similar behavior across languages and can be thought of as grammatically relevant universal categories humans use to conceptualize events. Following common terminology, we further refer to *general*, *predicate-independent* semantic roles as *thematic roles*. This level of granularity is used by VerbNet (Schuler, 2005) and, recently, VerbAtlas (Di Fabio et al., 2019).

Thematic roles' syntactic behavior depends on the presence of other thematic roles in the sentence: as our example above demonstrates, an `Instrument` can only take the subject position if the `Agent` is not present (3); and `Theme` can only become subject if both `Agent` and `Instrument` are not expressed (4-5). A widely used view to account for context dependency of semantic roles is *prominence*: given a syntactic prominence scale (e.g. *subject* ≺ *object*... ≺ *oblique*), one can assume that there exists a universal semantic prominence scale – a *thematic hierarchy* (TH) – which is homomorphic to the syntactic ranking (e.g. `Agent` ≺ `Instrument` ≺ `Theme`). Then, the top-ranking semantic argument in the hierarchy gets assigned to the highest available syntactic position, the second-ranking one gets the second-highest position, etc. While operating on thematic role level, VerbNet does not aim to provide a thematic hierarchy for its inventory[1].

Numerous thematic hierarchies have been proposed in linguistic literature (Rappaport Hovav and Levin, 2007). However, these proposals come from varying theoretical backgrounds, are based on different syntactic formalisms and operate with different role inventories. Most of the proposed hierarchies are justified via basic (often synthetic) language examples, aiming to verify a certain theory cross-lingually rather than to describe language use in a compact way. While no consensus on universally applicable thematic hierarchy could be reached, the notion of isomorphism between semantic and syntactic prominence is generally accepted. This led to several proposals which do not postulate the existence of a single hierarchy while still taking prominence into account. Pointing at the issues associated with atomic, categorical role inventories, Dowty (1991) suggests replacing them with Proto-Agent and Proto-Patient roles, each associated with a range of properties that determine the prominence of the corresponding semantic argument. Similarly, while generally following a predicate-specific categorical argument labeling scheme, PropBank uses an index-based approach where `Arg0` corresponds to Proto-Agent and `Arg1` to Proto-Patient; the rest of the core roles `Arg2-5` are underspecified with respect to prominence[2].

[1] Thematic hierarchy that ranks semantic roles by prominence is not to be confused with tree-like hierarchies used to organize role sets in (Bonial et al., 2011; Mújdricza-Maydt et al., 2016).

[2] Although generally Arg2-5 in PropBank are ordered according to their prominence (Martha Palmer, private correspondence).

Modern pre-trained contextualized language models are capable of generating coherent text, suggesting that they capture some notion of grammaticality. On the other hand, pre-trained encoders like BERT do not perform well on specialized grammaticality benchmarks (Warstadt et al., 2019) and struggle with predication-level phenomena like negation (Ettinger, 2020). From the perspective of probing, prominence is an attractive target as it is related to a variety of linguistic phenomena, incl. choice of subject, grammaticality and coreference. However, unlike local role assignments covered by the existing role probes, prominence only exists at a predication level: a semantic argument is not prominent *per se*, but is more, or less, prominent compared to another semantic argument of the same predicate. This calls for approaches to probing that can take the global structure and joint role assignments into account – or at least model prominence in a way that enables such a view.

4.1.2 Contributions

- We conduct an experiment in Semantic Proto-Role labeling and show that the information necessary to predict some of the Proto-Role properties can be located in the layer structure of pre-trained mBERT.
- We propose a regression-based framework for role labeling and use it to evaluate the capacity of the pre-trained mBERT model with respect to prominence.
- Our results suggest that mBERT implicitly encodes prominence; we show how regression-based role probes can interpolate prominence predictions based on partial data, and demonstrate that the probe has preference towards plausible role hierarchies.
- Finally, we propose an approach for corpus-driven thematic hierarchy induction. We operationalize the task and show that thematic hierarchies can be efficiently extracted from syntactically annotated corpora, resemble the proposals from the theoretical literature, and that the induced hierarchies can to some extent be applied cross-lingually.

4.2 Role Prominence

4.2.1 Categorical Role Inventories and SRL

Automatic semantic role labeling (SRL) is one of the core NLP tasks: given a sentence and a predicate (for example, a verb), the goal is to find the semantic arguments of this predicate in the sentence and assign them semantic roles. A predicate and its semantic arguments are referred to as *predication*. In SRL literature, this objective is often simplified to *“Finding out who did what to whom, where and when”*, making the task sound similar to fact extraction and eliminating any possible non-categorical interpretation of the role labeling objective. However, as we show below, such an interpretation is possible.

Role classifiers used in probing are mostly sourced from the categorical semantic role labeling corpora: PropBank, FrameNet fulltext corpus, OntoNotes (Pradhan and Xue, 2009) and others. The initial task formulation was proposed in the seminal work by Gildea and Jurafsky (2000); by now there exist multiple task setups for SRL, varying by formalism, predicate type (verbs, nouns), syntactic unit (dependency-based or constituents-based) and subtasks included (predicate identification, predicate disambiguation, argument identification and argument labeling). The relative importance of the subtasks highly depends on the formalism: for example, while predicate disambiguation might be of secondary importance for PropBank SRL, in FrameNet SRL the predicate sense determines the set of roles available for assignment, and predicate disambiguation turns out to be one bottleneck that might affect the overall frame-semantic labeler performance (Hartmann et al., 2017c). In terms of formalisms and languages, most work to date has been dedicated to English PropBank SRL, followed by frame-semantic parsing. Role probing focuses on the last stage of the SRL pipeline: argument labeling. Offering an alternative to traditional SRL, FitzGerald et al. (2018) cast categorical semantic role labeling as a question-answering task and demonstrate that it is possible to elicit SRL-like annotations in a crowdsourcing setup. Recent studies pave the path towards large-scale VerbNet-style semantic role labeling (Di Fabio et al., 2019; Gung and Palmer, 2021) and multilingual SRL based on parallel corpora (Daza and Frank, 2020).

From the modeling perspective, SRL systems have co-evolved with the rest of the natural language processing over the past decades, starting from the basic pipeline-based systems (Björkelund et al., 2009) and basic neural setups (Roth and Woodsend, 2014; FitzGerald et al., 2015). It was long believed that gold syntactic parses are a prerequisite to successful semantic role labeling; subsequent advances in end-to-end modeling made this requirement obsolete (He et al., 2017), although including syntax as a pruning mechanism (He et al., 2018) or multi-task auxiliary objective (Strubell et al., 2018) has constantly delivered performance gains over the syntax-agnostic models. Semantic role predictions must follow a set of theory-driven (e.g. each core role only appears once) and technical (e.g. a role must be licensed by the predicate sense) constraints; dedicated modeling approaches were proposed to take these into account, e.g. global optimization at inference time via integer linear programming as in Punyakanok et al. (2004), and others. In line with the trajectory of the field, task-specific end-to-end role labeling architectures were outperformed by a simple model built on top of a pre-trained BERT encoder (Shi and Lin, 2019). Conia and Navigli (2020) hold state of the art on multilingual dependency-based PropBank SRL with a similar model architecture.

Three major categorical role labeling formalisms are PropBank, FrameNet and VerbNet. Out of the three, only PropBank includes some notion of prominence in its role inventory: while the exact interpretation of the core argument labels `Arg0-5` is predicate-specific, `Arg0` corresponds to the Proto-Agent, while `Arg1` denotes Proto-Patient in Dowty's sense; for the rest of the core roles `Arg2-5` a prominence relationship holds generally, but is not guaranteed. Assuming such relationship makes it possible to treat PropBank core argument assignment as a ranking problem. We are aware of only one previous attempt to do so: a short paper by Sun et al. (2009) evaluates whether the thematic rank of the PropBank arguments can be detected based on syntactic parse information. They report role labeling experiments using a feature-based system, compare the ranking-based and categorical approaches to core argument labeling, and investigate the differences between two interpretations of ranking among PropBank roles: the one in which `Arg1` outranks `Arg2-4`, and the one in which `Arg1` is considered the least prominent argument among all. Although our setups are not comparable, we revisit some of the issues mentioned in their work in the modern context of probing.

4.2.2 Semantic Proto-Roles

The classical view on semantic roles as atomic, categorical entities is widely applicable and useful for explaining a range of syntactic and morphological phenomena, and the existence of few broad role-semantic categories like `Agent`, `Patient`, `Instrument` etc. is generally accepted. Creating a single universal role inventory, however, is challenging. The goal of designing a general-purpose role inventory is to be able to map every argument of every verb to a particular role, and at the same time to keep the inventory compact enough to generalize. Pointing out the contradictory nature of these requirements and the potential infeasibility of a universal role set, Dowty (1991) proposes to abandon the notion of a categorical role in favour of a prototype-based approach (Levin, 2019). Focusing on the subject choice of transitive verbs, he determines a set of ten entailment features such as "volitional involvement", "sentience", "change of state" and groups them into prototypical subject (Proto-Agent) and prototypical object (Proto-Patient) feature bundles.[3] The feature-based view on prominence avoids the problems with inventory fragmentation and overgeneralization that are characteristic to the category-based approach: the distinctions between roles are gradual, and the number of matching entailments for any given semantic argument determines its proximity to the prototypical Agent or Patient. A feature-based definition of semantic roles can accommodate the traditional categorical semantic roles, which are simply seen as frequently occuring feature combinations; and naturally incorporates the notion of prominence associated with the thematic hierarchy.

Despite its theoretical appeal, semantic proto-roles (SPR) have been operationalized in NLP only recently. Reisinger et al. (2015a) introduce the first large-scale corpus annotated with semantic proto-role properties. Demonstrating the practical advantage of property-based role definition, the authors were able to cast an otherwise expert-level role labeling task as a crowdsourcing task over a set of simple entailment questions. In the resulting dataset based on the Penn Treebank data (Marcus et al., 1993), each argument is assessed with respect to each of the Proto-Agent and -Patient properties on a 5-point

[3] In the discussion of Proto-Roles here and further, we use the terms *property* and *feature* interchangeably.

Likert scale. The authors report that some property combinations are indeed much more frequent, and compare the most prominent feature groups to the VerbNet roles of the corresponding arguments. In a subsequent work, White et al. (2016b) annotate a portion of the English Universal Dependencies treebank using a modified annotation protocol. Based on this data, several neural semantic proto-role labeling models have been proposed (Rudinger et al., 2018; Opitz and Frank, 2019). The data and annotation protocols were recently included into a larger Universal Decompositional Semantics toolkit (White et al., 2020).

Prominence is a general linguistic phenomenon, and semantic proto-roles offer one flexible way of modeling prominence. Given that the recent studies point at strong syntactic capabilities of pre-trained contextualized encoders (cf. Section 3.3.3), one would expect them to be able to represent prominence as well. To the best of our knowledge, the only existing probing study that involves proto-role semantics is (Tenney et al., 2019a); their results suggest that the information about proto-role features is evenly distributed across the model's layers. This is surprising, as the same study successfully localises the layers associated with syntactic parsing and traditional, categorical semantic role labeling – in line with our results reported in Section 3.3. To investigate this discrepancy, we expand our probing framework with a new set of regression-based proto-role labeling tasks.

4.3 Proto-roles in mBERT

For this experiment, we re-use the probing architecture introduced in Chapter 3, and expand the probing kit by incorporating eleven new semantic proto-role labeling tasks. Unlike Tenney et al. (2019a) who used a multi-label classification probe, we cast each of the proto-role properties as an independent regression task, use a shallower probe architecture and train separate scalar mixes for source and target tokens. Table 4.1 shows probing task examples. We populate the tasks from the original data by Reisinger et al. (2015a), resulting in a medium-sized probing dataset of approx. 5k sentences and 9.7k semantic arguments. The probe is trained for 20 epochs using the protocol described in Section 3.5.4. As before, we compare the probing performance to a randomly initialized mBERT model baseline.

Fitting a separate probing model for each proto-role property allows us to examine the probing performance in more detail. The results in Table 4.2 show that the performance varies by property, with some of the properties attaining reasonably high R^2 scores despite the simplicity of the probe architecture and the small dataset size. We observe that properties associated with Proto-Agent tend to perform better. The untrained mBERT baseline performs poorly which we attribute to the lack of data and the fine-grained semantic nature of the task.

Our fine-grained, property-level task design allows for more detailed insights into the layer utilization by the semantic proto-role probes (Figure 4.1). The

Proto-Agent	label	Proto-Patient	label
`instigation`	→ 2	`created`	→ 1
`volition`	→ 2	`destroyed`	→ 1
`awareness`	→ 5	`changes.possession`	→ 1
`sentient`	→ 5	`change.of.state`	→ 2
`change.of.location`	→ 3	`stationary`	→ 2
`exists.as.physical`	→ 5		

Table 4.1: Proto-role probing tasks with their corresponding proto-role feature values for the input *[She]$_{tgt}$ [saw]$_{src}$ a cat.*

property	R^2
(A) *instigation	0.68 (0.21)
(A) *volition	0.75 (0.11)
(A) *awareness	0.78 (0.09)
(A) *sentient	0.83 (0.07)
(A) *change.of.location	0.49 (0.04)
(A) *exists.as.physical	0.63 (0.03)
(P) *created	0.22 (0.01)
(P) *destroyed	0.11 (0.00)
(P) *changes.possession	0.26 (-0.01)
(P) *change.of.state	0.37 (0.01)
(P) *stationary	0.39 (0.05)

Table 4.2: Best dev R^2 for proto-role probing tasks over 20 epochs; A - Proto-Agent, P - Proto-Patient; Baseline in parentheses.

results indicate that while the layer utilization on the predicate side (`src`) shows no clear preference for particular layers (similar to the results obtained by Tenney et al. (2019a)), some of the proto-role features follow the pattern seen in the categorical role labeling and dependency parsing tasks for the argument tokens `tgt` (cf. Section 3.6). With few exceptions, we observe that the properties displaying that behavior are Proto-Agent properties; moreover, a close examination of the results on syntactic preference by Reisinger et al. (2015b, p. 483) reveals that these properties are also the ones with strong preference for the subject position, including the outlier case of `stationary` which in their data behaves like a Proto-Agent property. The correspondence is not strict, and we leave an in-depth investigation of the reasons behind these discrepancies for the future. While our results demonstrate that some aspects of prominence can be extracted from a pre-trained BERT encoder, the result is inconclusive, and we continue with an alternative, global regression-based setup for prominence probing.

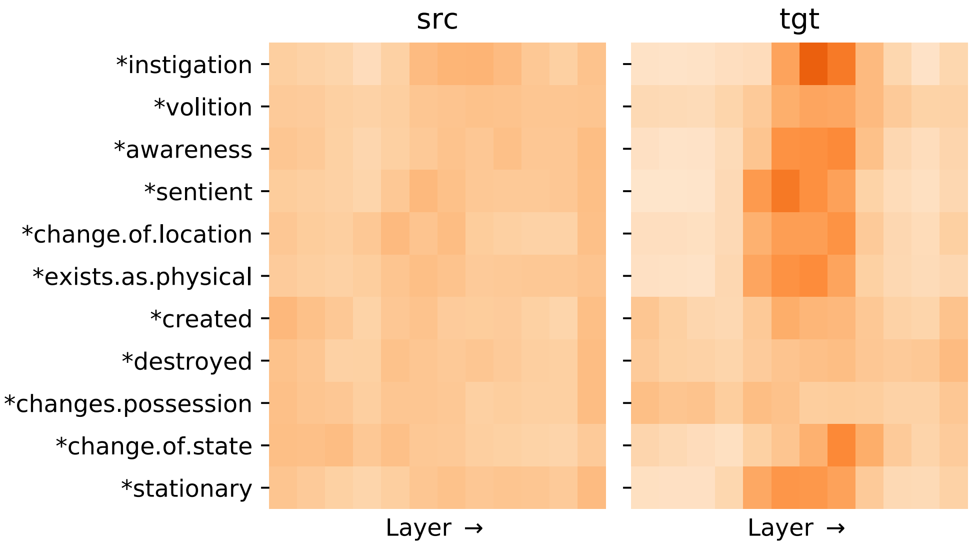


Figure 4.1: Layer utilization for semantic proto-role properties.

4.4 Prominence probing as role regression

4.4.1 Model

Like Sun et al. (2009), and similar to the setup used in edge probing, here we omit the argument identification stage of full SRL and solely focus on role labeling. The objective of the prominence-based regression role probe is to rank the semantic arguments in terms of their prominence. Formally, consider a generic categorical role inventory $R = \{r_1, r_2 ... r_n\}$. For simplicity, we assume that roles are general, i.e. not bound to a particular predicate. To map the categorical role set onto a continuous prominence scale, we define the global *prominence order* $o = r_1 \prec r_2 \prec ... \prec ... r_k$, where $r_i \prec r_j$ denotes that r_i is more prominent than r_j, similar to the traditional representation of thematic hierarchies. Using the prominence order, we then define a mapping $M : R \rightarrow \mathbb{R}$ that positions each categorical role on a linear scale such that for any $r_i \prec r_j$, $M(r_i) < M(r_j)$. As result, each categorical role is converted to a scalar value, independent of context.

Given a sentence s, predicate p and argument a, the regression-based role probe predicts the position of the argument on the prominence scale, such that the *relative* prominence of the arguments with respect to the predicate p is preserved. Figure 4.2 illustrates the difference between categorical and regression-based role labeling setups. Note that in addition to modeling a global relationship between roles by projecting them onto the shared linear scale, the prominence-based setting naturally satisfies the role uniqueness constraint as the arguments can be mapped onto the joint prominence scale. Moreover, eliminating the need for categorical labels allows us to interpolate the model predictions to the role labels never seen during training, and we explore this

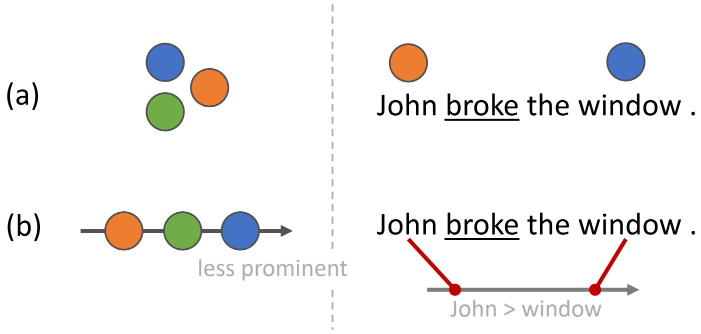


Figure 4.2: Role probing as role classification (a) and as prominence regression (b): a classifier in (a) assigns a categorical role, wile a regressor in (b) is only concerned with the relative placement of the semantic arguments on the prominence scale.

	train	dev
sentences	35.5k	1.2k
instances	166.8k	5.8k
Arg0	60.3k	2.0k
Arg1	83.5k	2.9k
Arg2	19.6k	0.7k
Arg3	3.4k	0.1k

Table 4.3: CoNLL-2009 prominence probing task statistics and label value distribution.

possibility below.

4.4.2 PropBank Instantiation

We instantiate our model with the PropBank formalism. Since here we only experiment with PropBank roles, we use the full CoNLL-2009 dataset instead of the SemLink-mapped CoNLL-2009SL that we used in Chapter 3. We focus on verbal predicates and core PropBank arguments, and discard the continuation (`C-`) and reference (`R-`) roles, as well as modifier roles (`AM-`). We further note that `Arg4` and `Arg5` are very rare in the data with less than 100 instances of `Arg4` and only three instances of `Arg5` in the development set of CoNLL-2009. To keep our setup compact and later experiments with optimal ranking in Section 4.4.5 feasible, we limit our scope to the roles `Arg0-3`. The statistics of the resulting dataset are shown in Table 4.3.

As mentioned before, PropBank generally assigns `Arg0` to Proto-Agent and `Arg1` to Proto-Patient roles. The prominence status of `Arg2-5` is underspecified; however, for the sake of an experiment we assume a linear mapping from PropBank roles to a prominence scale that maps the PropBank arguments to the integer positions corresponding to their role number. We revisit this assumption below.

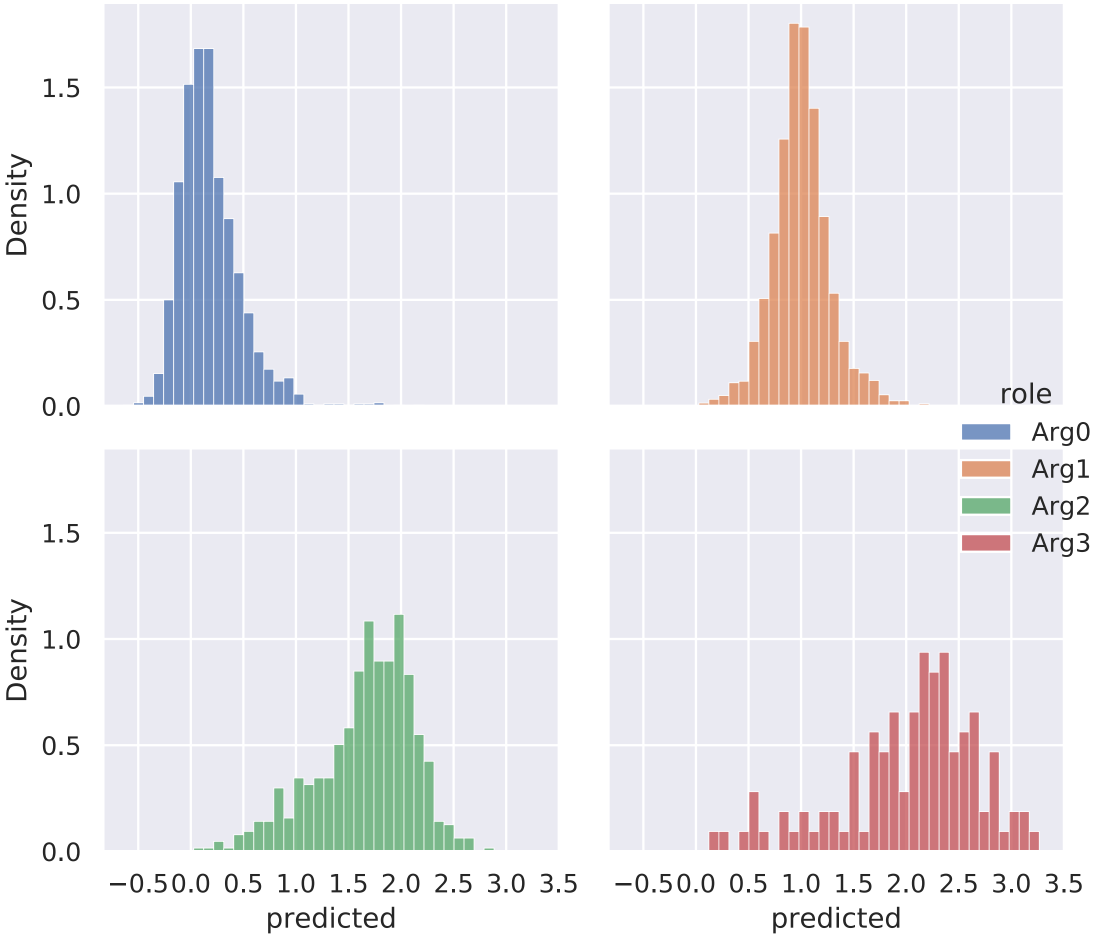


Figure 4.3: Prominence score distribution in the regression probe dev set predictions, by original PropBank role label.

We employ a regression model based on the probing architecture introduced in Chapter 3: given a sentence, a pre-trained frozen mBERT-base encoding of the source (predicate) and target (argument) token is used as input, followed by a trainable layer-wise scalar mix and a linear projection layer. While the previous regression probes reported in this work were evaluated via R^2; in case of prominence we are interested in mere *ranking* of the arguments and not in the absolute prominence values predicted by the model. Therefore we evaluate the models via rank accuracy by converting the prominence assessments for the roles to ranks within one predication.

4.4.3 General results

We first evaluate the output of the regression-based prominence model and compare it against a categorical PropBank role labeling probe. Both models

$\#(a)$	regression	classification	random init	reg. on `Arg0|3`
1	1.0	1.0	1.0	1.0
2	0.96	0.95	0.83	0.93
3	0.88	0.87	0.71	0.83
4	0.61	0.75	0.42	0.57

Table 4.4: Prominence probe performance for regression and classification probes, and a randomly initialized regression baseline, aggregated by the number of semantic arguments per predicate $\#(a)$. Rightmost column: probe performance for the interpolating regression probe.

are trained on the training split of CoNLL-2009 data and evaluated on the development split.

As a summary of the prominence predictions made by a trained probe, Figure 4.3 shows the distribution of the predicted values with respect to the gold PropBank categorical roles. We see that while the alignment is not perfect, the distributions of the predicted prominence generally capture the gold values obtained by mapping PropBank roles to a linear scale based on the role index.

Next we compare the performance of classification- and regression-based prominence probes. To do so, we convert both classification and regression predictions into ranks. An argument is considered to occupy the highest rank 1 among the other arguments of the same predicate if it has been labeled by a highest-ranking role or has been assigned the highest prominence score. Note that this evaluation criterion is more relaxed than strict role labeling, as it does not differentiate between the roles as long as the prominence relationship between them holds: for example, given two arguments, a model that predicts `Arg1` `Arg2` instead of `Arg0` `Arg2` would not be penalized, but a model that predicts `Arg0` `Arg1` instead of `Arg1` `Arg0` will. This relaxation is necessary to make regression- and classification-based model predictions comparable. We convert each role prediction to its rank and evaluate via rank assignment accuracy.

Table 4.4 presents the results. As rank accuracy always defaults to 1.0 if a proposition contains a single argument, we report the results separately for each argument set size. As we can see, the regression-based prominence model generally performs on par with the classification-based model, slightly outperforming it for 2- and 3-argument predications and underperforming for predications with 4 arguments. The regression probe baseline that uses a randomly initialized mBERT model shows a non-negligible level of performance, but is consistently worse than the classification and regression probes across all argument set sizes.

In line with our probing experiments, Figure 4.4 compares the scalar mixing weights of mBERT layers for both probe architectures. These results are not directly comparable to our observations in Chapter 3 due to the difference in data; however, we note that the classification-based PropBank probe trained on full CoNLL-2009 data generally attends to the same layers as the classification-based probe based on SemLink-mapped CoNLL-2009 that we

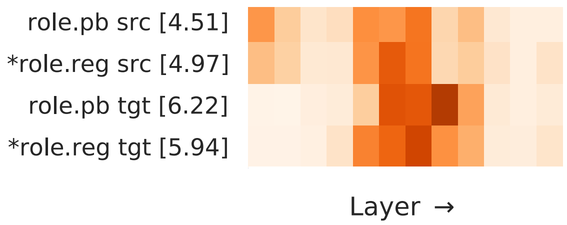


Figure 4.4: Layer utilization for classification (pb) and regression (reg) based prominence probes.

used in the earlier experiments. We attribute the differences to a larger dataset size and the higher frequency of predicate-speficic `Arg(2+)` roles. While not identical, classification- and regression-based probes attend to similar groups of layers in mBERT; this shows that the regression-based probe utilizes the information associated with role predictions, but the exact contribution of the individual layers involved in the prediction differs.

4.4.4 Interpolation

A regression-based approach to modeling role prominence naturally allows the model to predict prominence of the argument types it has not observed during training. To illustrate this property, we conduct an additional experiment with an *interpolated probe*, in which we filter out the roles corresponding to `Arg1` and `Arg2` from the training data and train the regression-based probe solely on `Arg0` and `Arg3` annotations. At the evaluation stage, we evaluate the ranking accuracy of the probe against the *full* role set on the development split, identical to the main evaluation setup.

We start by analyzing the predictions of the interpolating regression probe; as Figure 4.5 demonstrates, although the probe expectedly makes most confident predictions for `Arg0` and `Arg3` which were available during training, surprisingly, the predicted prominence values for `Arg2` and `Arg3` follow the order set by the prominence mapping M, suggesting that the probe indeed extracts prominence information from the mBERT model, albeit not perfectly. This is supported by the rank accuracy evaluation results presented in Table 4.4: although underperforming, the interpolated probe still maintains a reasonable level of rank accuracy across all argument set sizes and outperforms the baseline by a large margin. While the exact scalar prominence predictions deviate from the original mappings based on PropBank role index, the ranking of the arguments remains stable, suggesting that generalization takes place.

4.4.5 Optimal ranking

Unlike in categorical role probing where simply swapping two labels does not affect the results, regression-based probing depends on the definition of the role prominence order. Our results so far have been based on the assumption that

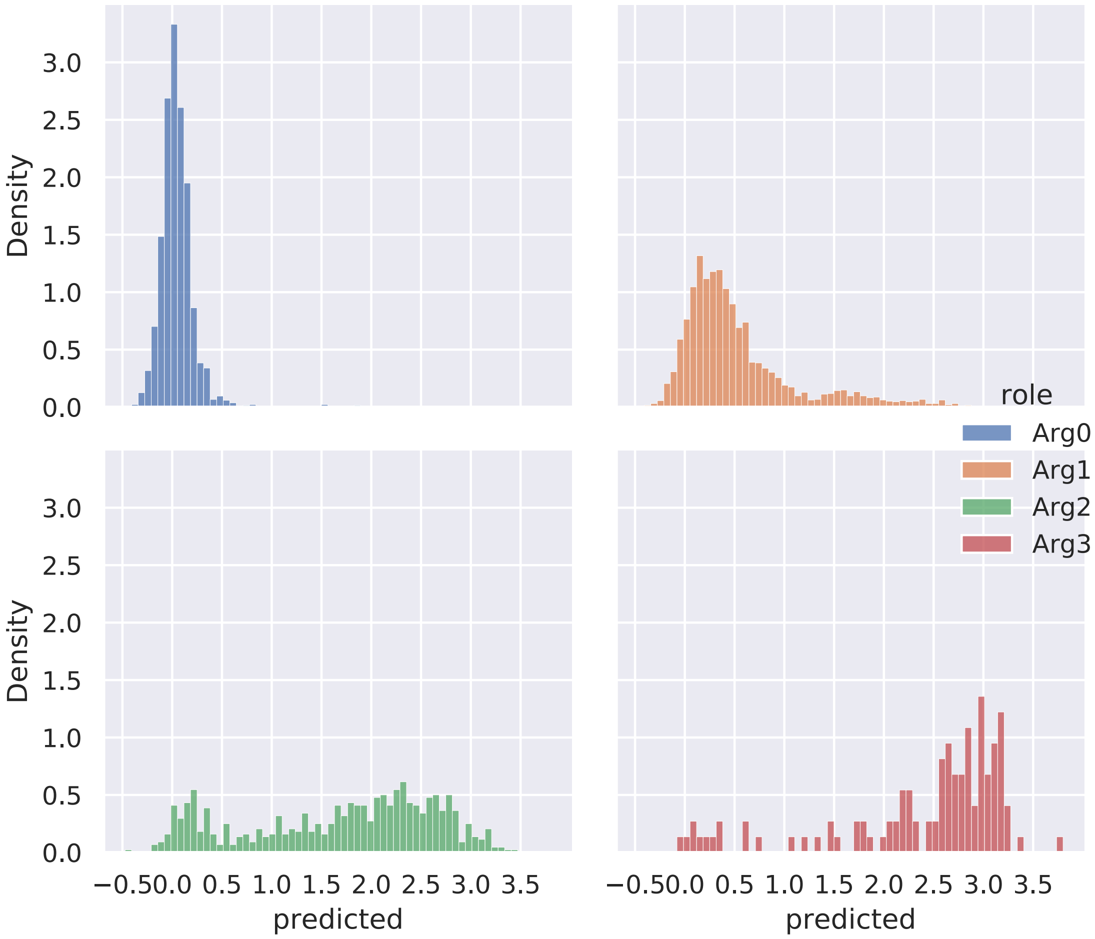


Figure 4.5: Interpolated regression probe predictions, by original PropBank role label. The probing model has only observed the roles `Arg0` and `Arg3` during training.

the optimal mapping of the PropBank roles to the prominence scale follows the PropBank role index, i.e. `Arg0` is mapped to $M(Arg0) = 0$ on the prominence scale, `Arg1` is mapped to $M(Arg1) = 1$ etc. However, this assumption is not strictly justified neither by the PropBank annotation guidelines, nor by linguistic theory: thematic hierarchy proposals vary as to whether the Patient role directly follows the Agent on the prominence scale or is situated at the very end of it (Rappaport Hovav and Levin, 2007; Sun et al., 2009). To investigate the viability of the chosen mapping and the impact of the mapping on prominence probing results, we conduct an exhaustive search over all possible rankings of the first four PropBank arguments. For readability, we use the shortcut `AX AY AZ` to denote a mapping $M(ArgX) = 1$, $M(ArgY) = 2$, $M(ArgZ) = 3$, etc. We generate all possible permutations of the roles `Arg0-3` and exclude the exact reverse duplicates. We then train a probe using the corresponding mapping M and evaluate it on the development set via rank

Arg order	$\#(a) = 2$	$\#(a) = 3$	$\#(a) = 4$
A2 A0 A1 A3	0.890	0.636	0.357
A0 A2 A1 A3	0.892	0.742	0.500
A1 A0 A3 A2	0.893	0.812	0.571
A0 A3 A2 A1	0.896	0.785	0.607
A2 A0 A3 A1	0.902	0.679	0.536
A1 A0 A2 A3	0.902	0.798	0.500
A1 A2 A0 A3	0.903	0.736	0.500
A0 A2 A3 A1	0.911	0.804	0.786
A2 A1 A0 A3	0.939	0.744	0.321
A0 A3 A1 A2	0.941	0.779	0.536
A0 A1 A3 A2	0.954	0.867	0.500
A0 A1 A2 A3	0.956	0.883	0.643

Table 4.5: Ranking accuracy depending on the prominence mapping of PropBank roles, sorted by 2-argument performance. $\#(a)$ denotes the number of semantic arguments per predicate.

accuracy. To keep the experiment feasible, we use a 5000-sentence sample for training.

The results in Table 4.5 show that the choice of the role ranking indeed has an effect on the regression-based prominence probing results. The rankings `A0 A1 A2 A3` and `A0 A1 A3 A2` result in the best fit for our regression-based probe, justifying our choice of M. Interestingly, while underperforming among 2-argument and 3-argument predicates, the alternative schemata that put the Patient at the end of the ranking perform well in the 4-argument subset. `A0 A2 A3 A1` performs best on 4-argument predicates, and its variation `A0 A3 A2 A1` is also among the best for this predicate arity. This preference for meaningful role prominence scales is another indication of prominence-related capabilities of the pre-trained mBERT model.

4.4.6 Discussion and Outlook

Existing work in probing has only considered semantic roles in isolation. We have introduced a regression-based prominence probe that investigates the higher-level grammatical capabilities of pre-trained contextualized encoders by mapping categorical roles onto a linear prominence scale. We have shown that the probe makes meaningful predictions, has interpolation capacity to predict prominence for previously unseen arguments, and prefers argument rankings consistent with the task logic.

Our probing model can accommodate any categorical role-semantic formalism that assigns global, predicate-independent role labels. The prominence mapping function is flexible and allows us to position “similarly prominent” categorical roles closer in the space, incl. rank equivalence of roles that do not co-occur; for example, an experiment variation equivalent to the one reported in (Sun et al., 2009) would treat `Arg2-4` as equally prominent. A similar setting that contextualizes the model predictions by mapping them on a linear scale can be applied to probe for other sentence-level phenomena that naturally allow such mapping, e.g. the choice of antecedent for pronouns (*John*$_1$ *gave Peter*$_2$ **his coat*) or discourse salience (*The president*$_1$ *visited the hospital*$_2$ vs *The hospital*$_1$ *was visited by the president*$_2$). While our previous approach to proto-role probing was cast as a feature-level prediction problem, semantic arguments can be naturally mapped onto a prominence scale based on their proto-role feature values.

Our pilot study on prominence probing with PropBank has several limitations that can be addressed in future work. We have only considered the core roles `Arg0-3`, however, the experiment can be easily extended to include the less frequent `Arg4-5` and the non-core ArgM tags, and take into account the reference and continuation roles. The prominence mapping of the PropBank roles can also be experimented with: while we assumed a simple monotonous mapping based on the role index, one could consider alternative allocations. While our interpolation experiment has shown promising results, alternative interpolation targets could be explored, e.g. to evaluate the probe's ability to learn prominence of unseen semantic roles solely based on `Arg0` and `Arg1` inputs. Finally, our results suggest that a regression-based formulation could be an interesting alternative to full categorical semantic role labeling, although the treatment of modifier roles `AM-` and selectional restrictions should be taken into consideration in this case.

From the methodological perspective, as with other structural probing approaches, both proto-role and regression-based role prominence probes should be supplemented by a behavioral probing setup as discussed earlier in Section 3.3.2, e.g. based on the existing grammaticality benchmarks (Warstadt et al., 2019; Ribeiro et al., 2020). The recently proposed amnesic probing approach (Elazar et al., 2021) that measures the effect of "masking out" certain information within the pre-trained model on the task performance is a promising way to bridge the gap between behavioral and structural probing and could be applied to the case of prominence probing as well.

From a linguistics perspective, one natural candidate for the expansion of our study would be VerbNet with its general-purpose thematic role inventory. However, unlike PropBank, VerbNet role labels per se do not suggest any hierarchical relationship, and VerbNet inventory size makes exhaustive search prohibitively expensive. Although several thematic role hierarchies have been proposed in literature, they are not directly compatible with the VerbNet thematic role set, and there exists no consensus over the feasibility of a general-purpose thematic hierarchy. Contributing to the operationalization of thematic hierarchies in NLP, in the next Section we propose a framework for inducing such hierarchies from VerbNet role annotations and syntax-labeled corpus data.

4.5 Thematic hierarchy induction

To recap, thematic hierarchies (TH) assume that given a syntactic hierarchy (e.g. *subject* $\prec$[4] *object* $\prec$ *oblique*) semantic roles can be ranked in a way that higher ranked roles take higher-ranked syntactic positions. One example of phenomena captured by THs is the choice of subject: given a thematic hierarchy `Agent` $\prec$ ... $\prec$ `Instrument`, an `Instrument` can only become subject if the `Agent` is not present, e.g. *"[John]$_{Ag}$ broke the window with a [hammer]$_{In}$"* → *"A [hammer]$_{In}$ broke the window"*. THs is a compact delexicalized way of explaining the interactions between the roles of the semantic arguments within one predication; however, while the existence of a general prominence relationship is accepted in linguistics, defining a universal role inventory and a universal thematic hierarchy has not been successful so far (Levin, 2019). Here, we offer a computational perspective on this question: we suggest a framework for corpus-driven thematic hierarchy induction and evaluate it qualitatively and quantitatively on English and German.

4.5.1 Thematic roles in NLP

THs have received considerable attention in linguistic literature, but have so far been impractical for use in NLP due to incompatibility and limited scope of the existing hierarchies. As a first step towards including THs into the NLP tool inventory, we suggest an empirical framework for inducing THs from role-annotated corpora. Since VerbNet (Schuler, 2005) was the only SRL framework that operates with thematic roles available at the time of this study, we chose it as our basis and performed experiments on the PropBank corpus (Palmer et al., 2005b) enriched with VerbNet role labels via SemLink (Bonial et al., 2013). Few studies have considered the VerbNet level of granularity in the past: Zapirain et al. (2008) compare PropBank and VerbNet performance using a simple SRL system and conclude that PropBank labels generally perform better; however, they do not use any additional modeling possibilities offered by VerbNet's general, predicate-independent role set. Loper et al. (2007) show that replacing verb-specific PropBank roles `Arg2-4` with the corresponding VerbNet roles improves the SRL performance. Merlo and van der Plas (2009) report a statistical analysis of PropBank and VerbNet annotations and conclude that while PropBank role inventory better correlates with syntax and is therefore easier to learn, VerbNet thematic roles are more informative and better generalize to new verb instances. A comparison on German data by Hartmann et al. (2017a) positions the VerbNet inventory above FrameNet and below PropBank in terms of complexity and generalization capabilities; however, the experiment is again based on the *mateplus* system (Roth and Woodsend, 2014) designed with PropBank generalization level in mind. Reisinger et al. (2015a) investigate the alignment between Dowty-style role properties and VerbNet thematic roles and show that VerbNet `Agents` tend to bear Dowty's `instigated`, `awareness` and `volitional` properties,

[4] We use $\prec$ for rank precedence, and / for ties

while Themes are more likely to change posession, change state, etc. After our study, the VerbAtlas project (Di Fabio et al., 2019) has released mappings for CoNLL-2009 data using a VerbNet-inspired thematic role inventory; while here we use our own projections driven by SemLink, our methodology is equally applicable to the VerbAtlas-style thematic roles.

4.5.2 Thematic hierarchies and syntactic formalisms

Numerous THs have been proposed in the linguistic literature, e.g. Agent $\prec$ Instrument $\prec$ Theme (Fillmore, 1968); see (Levin and Rappaport Hovav, 2005) for an overview. These hierarchies are rarely applicable for NLP since they originate from different theoretical backgrounds and are usually focused on a narrow set of linguistic phenomena (e.g. subject selection), aiming to provide a cross-linguistically valid hierarchy based on a set of manually constructed examples. In contrast, our approach is data-driven and aims to describe the general syntactic behavior of thematic roles. While an optimal TH that would successfully describe semantic roles' behavior across languages might not exist (and would imply the existence of a universal role inventory and grammar), our evidence suggests that this concept is at least partially applicable.

To the best of our knowledge, there exists no prior work explicitly aiming at discovering thematic hierarchies in corpora. However, the hierarchy-related effects are reported in some studies. For example, White et al. (2017) observe on a reduced role set that VerbNet roles disprefer the violations of thematic/syntactic hierarchy alignment. Sun et al. (2009) experiment on thematic rank prediction for PropBank Arg0 and Arg1, but extend their analysis neither to VerbNet thematic roles, nor to the PropBank Arg2-5.

Cross-lingual applicability has traditionally been a strong component in semantic role theory, and universality is one of the common desiderata for a thematic hierarchy. This, however, implies the existence of a universal syntactic prominence scale. From the NLP perspective, the closest to universal syntactic representation for which automatic parsers are available is the Universal Dependencies (UD) representation (Nivre et al., 2016), and we make an effort to ground our study in UD syntax for English. Since neither gold UD annotations, nor a deterministic converter were available at the time of the study, for German we use the TIGER dependency syntax representation (Brants et al., 2002).

4.5.3 Hierarchical linking model

We suggest a simple model to describe the interface between the syntactic and thematic rankings. An SRL corpus can be seen as a collection of sentences with corresponding predications, where each predication has a target (e.g. verb) and a set of arguments labeled with semantic roles.

Let $a_1 ... a_n \in A$ be the set of arguments in the predication p; $r(a_i)$ be the role label of the argument a_i, and $d(a_i)$ be the path between the predicate and the argument in the dependency parse tree of the sentence. A *syntactic ranker* S provides a syntactic rank $s_i = S(d(a_i))$ for each argument a_i in A based on the path, and a *thematic ranker* T provides a thematic rank $t_i = T(r(a_i))$ based on the argument's role – similar to the prominence mapping function that we used in our regression-based experiments Section 4.4. For each pair of arguments (a_i, a_j) we expect their syntactic ranks to align with their thematic ranks, i.e.

$$\forall i \neq j : sign(t_i - t_j) = sign(s_i - s_j)$$

The model per se does not imply the existence of a global ranking and allows flexible ranker definition. It allows ties in both syntactic and thematic rankings.

We use accuracy to assess how well a given syntactic-semantic ranker pair reflects the actual argument ranks found in data. Given a set of test predications $p_1, p_2 ... p_k \in P$ with the argument sets $A^1, A^2 ... A^k$, we measure the correspondence between syntactic and semantic ranking over the argument pairs (a_i^k, a_j^k) via accuracy defined as

$$\frac{\#(sign(t_i^k - t_j^k) = sign(s_i^k - s_j^k))}{\#total\_pairs}$$

To avoid the majority class bias, we measure accuracy for each role pair and use macro-averaged accuracy over pairs as the final score. A straightforward alternative to our evaluation metric would be the Kendall rank correlation coefficient, which, based on our preliminary experiments, tends to overemphasize the performance on the most frequent role pairs.

4.5.4 Induction Strategies

We investigate several thematic ranking strategies. As a running example for illustration, we use a small role set: `Agent` (`Ag`), `Patient` (`Pa`), `Instrument` (`In`), `Theme` (`Th`) and `Value` (`Va`). For now we assume the following syntactic hierarchy: *subj* $\prec$ *iobj* $\prec$ *nmod* $\prec$ *obj* $\prec$ *other.*

Local ranker The simplest way to model role ranking is to extract the average syntactic rank for each role based on the data, and then, given a test pair, assign ranks based on average syntactic rank.

role	Ag	Pa	In	Th	Va
$mean(s)$	1.01	2.58	1.72	3.95	3.74

Table 4.6: Mean syntactic rank per role (1-5)

Pairwise ranker Given that roles often strongly prefer a certain syntactic position (also see (White et al., 2016a)), local ranking is a reasonable baseline strategy. However, it fails to account for the context dependency of thematic

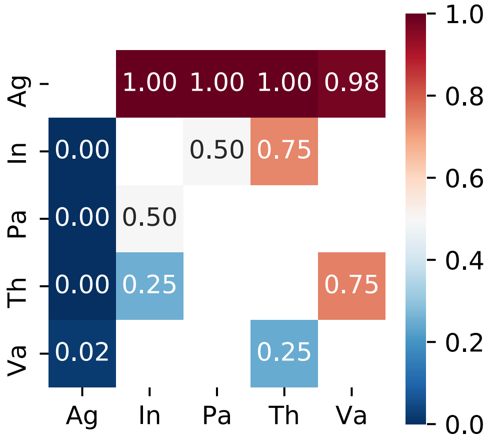


Figure 4.6: Preference matrix for the running example; the cell value corresponds to the number of times role on y axis has syntactically outranked the role on the x axis in the corpus, according to the pre-defined syntactic hierarchy.

roles' syntactic realization. The next step is to construct a *pairwise preference matrix*: for each pair of roles encountered in training data we calculate the proportion of times role r_i receives a higher syntactic rank than role r_j. For our role set, this results in the matrix shown on Figure 4.6. The preference matrix, for example, shows that `Agent` clearly dominates all the roles, `Instrument` outranks the `Theme`, and `Value` is below `Theme`.

Global ranker The pairwise ranking approach takes context into account. However, some role pairs only co-occur rarely. In such cases, no pairwise ranking information is available to the model. Finding a global TH based on pairwise preferences is an example of a rank aggregation problem which can be solved via constrained integer linear programming (ILP) optimization on a *preference graph* (Conitzer et al., 2006). We represent the pairwise preference matrix as a graph $G = (v, e)$ where each vertex v represents a role, the edge weight is the preference strength measured as $\#(r_i \prec r_j)/\#(r_i, r_j)$. The edge direction is from higher- to lower-ranking role. If we assume a global ordering of the roles, we can induce the global ranking via transitivity relations. For example (Figure 4.7), `Instrument` never appears in the same sentence as `Value` in our training data; however, by transitivity via `Theme` we can assume that `Instrument` ranks over `Value`.

Given the preference graph $G = (v, e)$, let w_{ij} be the weight of the edge between v_i and v_j. Let $x_{ij} \in 0, 1$ denote that we rank node v_i above v_j. The goal is then to maximize $\sum_{i,j} x_{ij} w_{ij}$ subject to two groups of constraints. First, we prohibit two nodes to rank above each other, but allow ties, by enforcing $\forall_{i,j} : x_{ij} + x_{ji} \leq 1$. Second, we enforce transitivity, i.e. if r_i is ranked above r_j, and r_j is ranked above r_k, then r_i must be ranked above r_k, formally

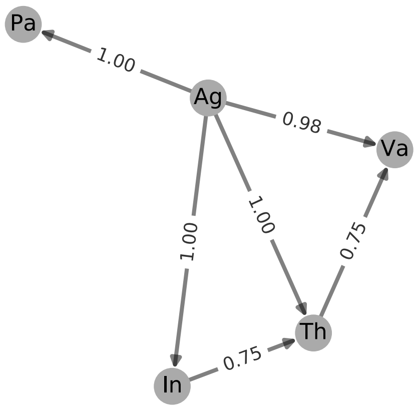


Figure 4.7: Preference graph corresponding to the preference matrix in Figure 4.6; nodes denote semantic roles, edges show the degree of preference as observed in the corpus, from less to more prominent role.

$\forall_{i,j,k}, i \neq j \neq k : x_{ij} + x_{jk} - x_{ik} \leq 1$. We solve the ILP problem using the off-the-shelf *pulp* optimizer (Mitchell et al., 2011).

For our restricted example, the optimization produces the following global hierarchy: `Ag` $\prec$ `In` $\prec$ `Th` $\prec$ `Va`/`Pa`. This hierarchy ranks `Instrument` above `Value` by transitivity, however, in case of `Patient` and `Value` no preference can be inferred from the graph, so they receive the same thematic rank.

4.5.5 Setup

Datasets and restrictions

For our experiments on English, we use SemLink (Bonial et al., 2013), a manually constructed resource that enriches PropBank (Palmer et al., 2005b) semantic role annotations for Penn Treebank with VerbNet (Schuler, 2005) thematic role labels. We use the Universal Dependencies converter (Schuster and Manning, 2016) to transform original PropBank syntactic annotation into UD v. 1.0. PropBank semantic role annotation and the corresponding SemLink reference are constituents-based. However, UD is a dependency formalism, and we employ a number of heuristics to align original PropBank annotations with the CoNLL-2009 datasets (Hajič et al., 2009) to recover the head node positions. We employ additional transformations, filtering out the predications in which not all PropBank core roles got aligned to the VerbNet thematic roles.

For German, we use the SR3de dataset (Mújdricza-Maydt et al., 2016; Hartmann et al., 2017a) which explicitly provides VerbNet annotations on top of the SALSA corpus (Burchardt et al., 2006). At the moment of the study, there

dataset	#sent	#tok	#pred	#arg
EN (PropBank→SemLink)				
train	16 603	446 641	21 276	44 333
test	1 031	27 751	1 336	2 761
dev	550	15 157	684	1422
DE (SR3de VerbNet)				
train	898	20 277	905	1 992
test	240	4 738	245	532
dev	117	2 429	119	266

Table 4.7: Dataset statistics

existed no gold UD annotations for SALSA, and we use its default TIGER syntactic formalism (Brants et al., 2002) in our experiments.

Following previous work, we impose certain restrictions on our data. Since thematic roles in both VerbNet and SR3de are only defined for verbal predicates, we restrict the scope of our study to verbs. We only consider direct dependents of the verbs in active voice, and since having access to the full argument set is important to study context dependency, we only consider the predications where all arguments are direct dependents of the verb in the Universal Dependencies tree. Since we are interested in relative ranking, only predications that contain more than one semantic argument are considered in the study.

Dataset statistics for English and German (after filtering) are summarized in Table 4.7. In all experiments, we induce a TH and related statistics from the training data and evaluate it on the test data, using the split from the CoNLL-2009 shared task. Note that the resulting data differs from the VerbNet-enriched CoNLL2009-SL used for probing in Chapter 3: while our probing studies operate with native syntactic formalisms, here we use Universal Dependencies. Furthermore, the role probe is based on the arguments mapped to VerbNet *and* FrameNet in SemLink, but here we are only interested in VerbNet thematic roles (and can thereby retain more data); to keep the study tractable, we perform additional syntax-based filtering of the data here, but not in the probing study.

Syntactic ranker

For simplicity in this study we only experiment with two syntactic rankers per language. A common syntactic prominence scale assumed in linguistic literature is *subject* ≺ *object* ≺ *indirect object* ≺ *oblique*. This scale has to be adapted to the UD and TIGER labeling schemes. For each language we evaluate two syntactic rankings: one that positions *objects* above *indirect objects* and *obliques*, and one that positions *objects* below.

For English, we rank the UD syntactic relations as follows (**SE1**): *nsubj* / *csubj* ≺ *iobj* ≺ *nmod* ≺ *ccomp* / *dobj* ≺ *other*; where *nmod* corresponds to oblique and *other* is used for any other syntactic relation. An alternative ranking

	synt	glob	pair	loc	RND	UB
EN	SE1	.869	.887	.867	.509	.927
EN	SE2	**.930**	**.929**	**.913**	.500	**.932**
DE	SD1	.655	.726	.637	.471	.818
DE	SD2	**.790**	**.820**	**.820**	.456	**.920**

Table 4.8: Thematic ranker evaluation for global (`glob`), pairwise (`pair`) and local (`loc`) rankers, as well as random ranker (`RND`) and upper bound (`UB`); bold – best result over syntactic rankers SE1/2 and SD1/2, underlined – best result over thematic rankers.

positions *dobj* directly after the subject (**SE2**): *nsubj* / *csubj* $\prec$ *ccomp* / *dobj* $\prec$ *iobj* $\prec$ *nmod* $\prec$ *other.*

For German, the following ranking of TIGER syntactic relations is employed (**SD1**): *SB* $\prec$ *DA* $\prec$ *OP* / *MO* / *OG*/ *OC* $\prec$ *OA* / *OA2* / *CVC* $\prec$ *other*; where *SB* is the subject, *DA* is dative object, *OP* / *MO* / *OG* / *OC* correspond to oblique relations, and *OA* / *OA2* / *CVC* to direct object relations (see (Brants et al., 2002) for detailed description). Similarly, we evaluate the performance of the ranking that positions the direct object after the subject (**SD2**): *SB* $\prec$ *OA* / *OA2* / *CVC* $\prec$ *DA* $\prec$ *OP* / *MO* / *OG* / *OC* $\prec$ *other.*

Bounds. We construct the *upper bound* for the hierarchy induction by evaluating a global ranker trained on the test dataset. The upper bound reflects the data properties, as well as the maximal alignment accuracy that can be achieved with the selected syntactic ranker. The **lower bound** is constructed by evaluating 100 random thematic rankers which rank roles according to a random (but consistent) hierarchy, and averaging the result.

Data utilization. To evaluate how effectively the proposed rankers use the training data, we conduct a series of experiments with reduced dataset sizes using the following protocol. The training dataset is shuffled and split into $n = 100$ slices. A ranker is consecutively trained on the first $m \in 1..n$ slices and evaluated against the full test dataset. The procedure is repeated $k = 100$ times to eliminate the effect of data order, and the results per slice are averaged.

4.5.6 General Accuracy and Syntactic Ranker

To get an overall impression of the ranking quality, we first compare the performance of thematic rankers with respect to syntactic rankers and available datasets. The results of this comparison are summarized in Table 4.8 and show that syntactic rankers positioning the object second in the hierarchy (SE2 and SD2) lead to better alignment on both datasets and have a higher upper bound. We report the results on these rankers for the rest of the chapter.

EN	Agent $\prec$ Cause/Instrument/Experiencer $\prec$ Pivot $\prec$ Theme $\prec$ Patient $\prec$ Material/Source/Asset $\prec$ Product $\prec$ Recipient/Beneficiary/Destination/Location $\prec$ Value/Stimulus/Topic/Result/Predicate/Goal/InitialLocation/Attribute/Extent
DE	Agent $\prec$ Experiencer $\prec$ Stimulus/Pivot $\prec$ Cause $\prec$ Theme $\prec$ Patient $\prec$ Topic $\prec$ Instrument $\prec$ Beneficiary/InitialLocation $\prec$ Result $\prec$ Product/Goal $\prec$ Destination/Attribute $\prec$ Recipient $\prec$ Value/Time/CoAgent/Locus/Manner/Source/Trajectory/Location/Duration/Path/Extent

Table 4.9: Induced hierarchies

For English the global hierarchy-based ranker approaches the upper bound, closely followed by the pairwise ranker. The accuracy on German data is lower and the pairwise and local rankers outperform the global hierarchy-based ranker. We revisit this observation later.

4.5.7 Qualitative analysis

The result of hierarchy induction is a global ranking of thematic roles. Table 4.9 shows full rankings extracted for English and German data. While some correspondence to the hierarchies proposed in literature is evident (e.g. for English `Agent` $\prec$ `Instrument` $\prec$ `Theme`, similar to (Fillmore, 1968)), a direct comparison is impossible due to the differences in role definitions and underlying syntactic formalisms. Notice the high number of ties: some roles never co-occur (either by chance or by design) or occur in the same syntactic rank (e.g. *oblique*) so there is no evidence for preference even if we enforce transitivity.

4.5.8 Cross-lingual hierarchy induction

The induced hierarchies for English and German bear certain similarities, which raises the question of cross-lingual applicability of the hierarchies. This analysis is only possible because the VerbNet and SR3de role inventories are mostly compatible with few exceptions (Mújdricza-Maydt et al., 2016). Table 4.10 contrasts the performance of THs induced from English and German training data, and evaluated on German and English test data respectively. While the cross-lingual performance is expectedly lower than the monolingual performance, it outperforms the random baseline by a large margin, suggesting the potential for cross-lingual hierarchy induction.

4.5.9 Data utilization

One can assume that constructing a global hierarchy should require less training data due to the effective utilisation of transitivity. We evaluate this assumption empirically. Figure 4.8 reports the performance of rankers with ac-

	EN-test	DE-test
UB	.932	.920
EN-train	.930	.787
DE-train	.852	.790
RND	.500	.456

Table 4.10: Cross-lingual evaluation, global ranker; random ranker (RND) and upper bound (UB) for comparison.

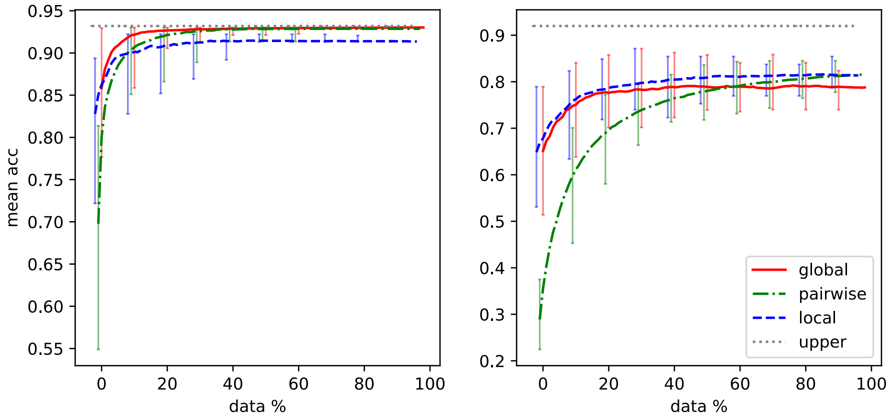


Figure 4.8: Data utilization for English (left) and German (right) along with max/min values

cess to different amounts of training data for English and German. The results on English data show that global hierarchy-based ranker effectively utilizes the training data and can be trained using just fractions of the original training dataset.

The accuracy measurements on German are less conclusive: the local ranker generally performs best and learns fastest. We attribute this to the fact that filtered SR3de is an order of magnitude smaller than our mapped and filtered PropBank-SemLink dataset. For pairwise and global rankers as many role pairs as possible should be observed at least once to establish the pairwise preference. This holds for our mapped PropBank-SemLink (all role pairs from test data seen at least once after observing 20% of the training data, on average), however, for filtered SR3de, even given the full training data, only 83% of role pairs from the test set have been seen at least once.

4.5.10 Error analysis

Our evaluation procedure allows detailed insights into the performance of the models. To illustrate, we extract the role pairs from English and German data with ranking accuracy below 1.0.

Table 4.11 lists the ranking inconsistencies produced by the global ranker for English. We can see that false ranking can be caused by the lack of training

Role pair	score	#(train)
Recipient - Topic	0.35	338
Source - Theme	0.46	246
Location - Theme	0.53	400
Material - Product	0.67	29
Result - Theme	0.67	30
Experiencer - Stimulus	0.74	922
Destination - Theme	0.86	401
Instrument - Theme	0.88	110
Recipient - Theme	0.89	419
Attribute - Experiencer	0.90	166

Table 4.11: Global ranker accuracy, English

Role pair	score	#(train)
Attribute - Source	0.00	0
Beneficiary - Manner	0.00	0
Beneficiary - Value	0.00	1
Extent - Goal	0.00	2
Goal - Recipient	0.00	12
Instrument - Result	0.00	3
Locus - Topic	0.12	3
Recipient - Theme	0.40	26
Recipient - Topic	0.50	5
Pivot - Theme	0.67	57

Table 4.12: Global ranker accuracy, German

examples (e.g. `Material` vs. `Product`, `Theme` vs. `Result`). We also observe complications with positioning the `Theme` in the hierarchy. In many cases, the misalignment is due to non-standard use of thematic roles, e.g. `Location` as subject in `wsj_2322:7` [*the* ***delay***$_{Loc}$ *resulted* *from* ***difficulties***$_{Th}$]. Another common reason for false alignments is the syntactic ranker. For example, in `wsj_2372:1` [*the* ***Senate***$_{Ag}$ *voted* ***87-7***$_{Res}$ *to* ***approve***$_{Th}$*...*] the `Result` is connected to the predicate via an *advmod* relation, and `Theme` is *xcomp*, both ranked equally (*other*) by our syntactic ranker.

Error analysis on the much smaller German dataset (Table 4.12) reveals the sparsity-related issues: most of the role pairs that tend to get misaligned do not, or only rarely appear in the training data, heavily influencing the score. As on English data, many misalignments are due to simplicity of the syntactic ranker.

4.5.11 Discussion

Importance of the syntactic ranker. The choice of syntactic ranking has a drastic effect on the resulting TH and the alignment quality, even if only direct syntactic dependents and a limited set of relations are taken into account.

Realistically there might exist an arbitrary set of paths connecting arguments to predicates. UD as a syntactic formalism is also subject to change. Although we show that THs can be induced with an arbitrary dependency formalism, a cross-lingual UD-based study would be another extension to our work. Recent work in low-cost generation of parallel data for SRL makes such study feasible (Daza and Frank, 2020).

Data selection. We have demonstrated that THs can be induced from small portions of training data. The large discrepancy in the scores on the first data slices seen in Figure 4.8 suggests that some data instances are more informative for TH induction. This raises the question of whether it is possible to automatically select useful training instances, supported by the evidence from previous work in SRL (Peterson et al., 2014; Myers and Palmer, 2021). One obvious strategy would be to make sure that the hierarchy inducer is presented with as many distinct role pairs as early as possible. Approximating this objective in an unsupervised way would reduce the amount of data needed to induce a high-quality thematic hierarchy.

The need for a global hierarchy. Our results regarding the necessity of a global hierarchy which ranks *all* the roles are inconclusive. While global ranking reaches the best quality for English, on the German data pairwise and local ranking approaches perform best. Although we attribute the latter to sparsity, more German data would be needed to evaluate this hypothesis. In particular, this can be achieved by relaxing some of the constraints we impose on the data.

Future research directions. One can divide the potential future research directions for this work into two categories. From the resource perspective, our methodology can be applied to the new VerbAtlas resource that uses a set of VerbNet-like thematic roles to describe predicate semantics while providing higher coverage (Di Fabio et al., 2019); recent work on constructing parallel multilingual corpora for SRL (Daza and Frank, 2020) and successful reports of projecting SRL annotations onto Universal Dependencies corpora (Akbik et al., 2015) bear potential for truly multilingual thematic hierarchy induction. From the methodological perspective, the weak point of the syntax-based thematic hierarchy induction model is its reliance on a pre-defined syntactic ranking. Recent advances in probing and the syntactic capabilities of modern pre-trained encoders (Chi et al., 2020) might mitigate this issue: an *explicit* syntactic ranking is not necessary for our approach to function, as long as relative syntactic prominence can be measured between two semantic arguments in any other way.

4.6 Chapter summary

While most work in NLP treats role labeling as a classification task, in this chapter we have explored an alternative, prominence-based view on role labeling. Continuing with layer probing as our main methodological framework, we have analyzed the proto-role modeling capabilities of the pre-trained mBERT model and demonstrated that some of the proto-role properties that determine prominence can be extracted from and localized in the model. We have presented a novel prominence probing model and reported extensive experiments on its instantiation based on widely used categorical PropBank roles. Finally, we have presented a corpus-driven approach to thematic hierarchy induction from linguistically annotated corpora and evaluated it for English and German. A discussion of hierarchy induction results, limitations and future research directions concludes the chapter.

Chapter 5

Conclusion

Linguistic theories and conceptualisations have been instrumental to the development of natural language processing. The advancements of the past decade have shifted the primary NLP methodology to transfer learning via strong pre-trained encoder models. Given the lasting success of the last-generation Transformers across tasks, fine-grained, expert-curated linguistic feature sets of the past are unlikely to re-emerge in mainstream NLP. However, despite the impressive performance, even todays' best-performing pre-trained encoders suffer from the lack of transparency, biases and unpredictability. As NLP starts to find its way into real-world applications, it becomes crucial to have a clear picture of the capabilities and internal mechanics of these strong pre-trained models. Interestingly, the extremely high cost of training general-purpose Transformer encoders from scratch forces the NLP community to share and re-use existing models; this creates an additional incentive for understanding how the few popular pre-trained encoders work, as the insights from analysing one such model can be of use for a variety of applications and tasks.

A key feature of contextualized Transformer-based encoders like BERT and GPT is their wide transferability across tasks: the same BERT model can be easily fine-tuned to reach state-of-the-art performance on word sense disambiguation and on recognizing textual entailment. This wide applicability suggests that strong encoders capture some generally useful knowledge about the language during pre-training, and the line of research in probing aims to investigate what this knowledge is and how the pre-trained models represent it. This idea is neither new nor specific to the Transformer-based contextualized models, and the early studies in similarity benchmarking and resource-based evaluation of static word embeddings like word2vec and GloVe are methodologically very close to modern probing approaches.

A pre-trained representation or model is always probed *for something*: a probing study requires a phenomenon the model would be evaluated against. While it is possible to design and generate data for an arbitrary probe ("Does this sentence start with a consonant"), a principled probing framework grounded in theory would be more useful across applications and languages. Linguistics provides a convenient scaffolding for probing, as linguistic theories are designed to generalize, do not depend on the domain or application, aim for cross-lingual generality, and are often supplemented by manually crafted resources and annotated corpora. Indeed, linguistic annotations are increasingly used as gold standards in evaluation and probing of pre-trained representations.

Unlike simple surface-level phenomena like character count or word content, linguistic annotations are usually associated with a theory. In this thesis we have explored how the assumptions and conceptualizations made by underlying theory might influence the results of probing studies. In Chapter 2 we highlight the conceptual gap between static word embedding vocabularies and lexical resource entries, and show that reducing this gap via grammatical normalization leads to better alignment of word embedding spaces and lexical resources in a novel resource-based evaluation setup. In Chapter 3 we hypothesize that the choice of linguistic formalism might affect probing measurements for modern pre-trained encoders and demonstrate that this is the case. We extend the state-of-the-art layer probing framework and provide many additional insights into the effects of data size, dataset and formalism implementation on probing results. Finally, in Chapter 4 we investigate how alternative linguistic theories might be used to further refine our understanding of pre-trained models in a range of experiments dedicated to probing for prominence. We report experiments on probing for semantic proto-roles, and find evidence of proto-role information being localized within the pre-trained BERT model, contrary to the existing reports. We propose a regression-based semantic role probe that – contrary to the NLP tradition – treats role labeling as a regression task, and use it to investigate the prominence-modeling capabilities of BERT. A non-neural approach to inducing prominence hierarchies for thematic, VerbNet-style role sets concludes the thesis: we show that thematic hierarchies – popular in linguistics but so far not suitable for NLP – can be induced from syntactically annotated corpus data, resemble the proposals from the linguistic literature and to an extent apply cross-lingually.

As NLP increasingly finds its way into our everyday lives, the demand for interpretability and transparency becomes more and more apparent – not only from the perspective of ethics and real-world consequences, but also from the perspective of informed, iterative, hypothesis-driven development of the NLP models and applications. Although linguistic phenomena remain a source of generalization power and an application target on their own, a perhaps even more important role of linguistics in todays' NLP is to serve as a common, theoretically grounded meta-language for us to talk about what our end-task models do well, what they lack, and why. In this context it is crucial to remember that linguistics does not simply amount to tasks and datasets, as is common in NLP: linguistic material is always bundled with theoretical background assumptions – be it the unit of analysis, the strategy used to group words together in a resource, or the task architecture itself. For linguistics to serve well as a robust "interpretation language" for modern and future NLP, those assumptions need to be studied and their effects taken into account, and we hope that the work described in this thesis contributes to our better understanding of the role of linguistics in NLP interpretation and probing task design.

List of Figures

List of Tables

Bibliography

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa: 'A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches', in: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 19–27, Association for Computational Linguistics, Boulder, Colorado, June 2009, Online: https://www.aclweb.org/anthology/N09-1003.

Lars Ahrenberg: 'Converting an English-Swedish Parallel Treebank to Universal Dependencies', in: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 10–19, Uppsala University, Uppsala, Sweden, Uppsala, Sweden, August 2015, Online: https://www.aclweb.org/anthology/W15-2103.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf: 'FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP', in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019, Online: https://www.aclweb.org/anthology/N19-4010.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu: 'Generating High Quality Proposition Banks for Multilingual Semantic Role Labeling', in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 397–407, Association for Computational Linguistics, Beijing, China, July 2015, Online: https://www.aclweb.org/anthology/P15-1039.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein: 'A Diagnostic Study of Explainability Techniques for Text Classification', in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3256–3274, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.263.

Ben Athiwaratkun and Andrew Wilson: 'Multimodal Word Distributions', in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1645–1656, Association for Computational Linguistics, Vancouver, Canada, July 2017, Online: http://aclweb.org/anthology/P17-1151.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe: 'The Berkeley FrameNet Project', in: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, pp. 86–90, Association for Computational Linguistics, Stroudsburg, PA, USA, 1998.

Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi: 'What's in a Name? Are BERT Named Entity Representations just as Good for any other Name?', in: *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 205–214, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.repl4nlp-1.24.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski: 'Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors', in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 238–247, Association for Computational Linguistics, Baltimore, Maryland, June 2014, Online: https://www.aclweb.org/anthology/P14-1023.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick: 'Interpretability and Analysis in Neural NLP', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 1–5, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.acl-tutorials.1.

Yonatan Belinkov and James Glass: 'Analysis Methods in Neural Language Processing: A Survey', *Transactions of the Association for Computational Linguistics* 7: 49–72, March 2019, Online: https://www.aclweb.org/anthology/Q19-1004.

Iz Beltagy, Kyle Lo, and Arman Cohan: 'SciBERT: A Pretrained Language Model for Scientific Text', in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–3620, Association for Computational Linguistics, Hong Kong, China, November 2019, Online: https://www.aclweb.org/anthology/D19-1371.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin: 'A Neural Probabilistic Language Model', *Journal of Machine Learning Research* 3: 1137–1155, March 2003.

Michele Bevilacqua and Roberto Navigli: 'Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by

Incorporating Knowledge Graph Information', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2854–2864, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.acl-main.255.

Anders Björkelund, Love Hafdell, and Pierre Nugues: 'Multilingual Semantic Role Labeling', in: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pp. 43–48, Association for Computational Linguistics, Boulder, Colorado, June 2009, Online: https://www.aclweb.org/anthology/W09-1206.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov: 'Enriching Word Vectors with Subword Information', *Transactions of the Association for Computational Linguistics* 5: 135–146, 2017, Online: https://www.aclweb.org/anthology/Q17-1010.

Claire Bonial, William J. Corvey, Martha Palmer, Volha Petukhova, and Harry Bunt: 'A Hierarchical Unification of LIRICS and VerbNet Semantic Roles', *2011 IEEE Fifth International Conference on Semantic Computing* pp. 483–489, 2011.

Claire Bonial, Kevin Stowe, and Martha Palmer: 'Renewing and Revising SemLink', in: *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pp. 9–17, Association for Computational Linguistics, 2013, Online: http://www.aclweb.org/anthology/W13-5503.

Cristina Bosco, Manuela Sanguinetti, and Leonardo Lesmo: 'The Parallel-TUT: a multilingual and multiformat treebank', in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1932–1938, European Language Resources Association (ELRA), Istanbul, Turkey, May 2012, Online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/209_Paper.pdf.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler: 'New Protocols and Negative Results for Textual Entailment Data Collection', in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8203–8214, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.658.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith: 'The TIGER Treebank', in: *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pp. 24–41, 2002.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran: 'Distributional Semantics in Technicolor', in: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 136–145, Association for Computational Linguistics, Jeju Island, Korea, July 2012, Online: https://www.aclweb.org/anthology/P12-1015.

Elia Bruni, Nam Khanh Tran, and Marco Baroni: 'Multimodal Distributional Semantics', *Journal of Artificial Intelligence Research* 49 (1): 1–47, January 2014, Online: http://dl.acm.org/citation.cfm?id=2655713.2655714.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal: 'The SALSA corpus: A German corpus resource for lexical semantics', in: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pp. 969–974, European Language Resources Association (ELRA), 2006, Online: http://www.aclweb.org/anthology/L06-1195.

Jose Camacho-Collados and Roberto Navigli: 'BabelDomains: Large-Scale Domain Labeling of Lexical Resources', in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 223–228, Association for Computational Linguistics, Valencia, Spain, April 2017, Online: https://www.aclweb.org/anthology/E17-2036.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli: 'SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity', in: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 15–26, Association for Computational Linguistics, Vancouver, Canada, August 2017, Online: https://www.aclweb.org/anthology/S17-2002.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom: 'e-SNLI: Natural Language Inference with Natural Language Explanations', in S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.): *Advances in Neural Information Processing Systems*, Vol. 31, pp. 9539–9549, Curran Associates, Inc., 2018, Online: https://proceedings.neurips.cc/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf.

Richard Eckart de Castilho and Iryna Gurevych: 'A broad-coverage collection of portable NLP components for building shareable analysis pipelines', in: *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pp. 1–11, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, August 2014, Online: https://www.aclweb.org/anthology/W14-5201.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia: 'SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation', in: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Association for Computational Linguistics, Vancouver, Canada, August 2017, Online: https://www.aclweb.org/anthology/S17-2001.

Eugene Charniak, Curtis Hendrickson, Neil Jacobson, and Mike Perkowitz: 'Equations for Part-of-Speech Tagging', in: *Proceedings of the Eleventh Na-*

tional Conference on Artificial Intelligence (AAAI-03), p. 784–789, The MIT Press, Washington, DC, July 1999.

Stanley F. Chen and Joshua Goodman: 'An Empirical Study of Smoothing Techniques for Language Modeling', in: *34th Annual Meeting of the Association for Computational Linguistics*, pp. 310–318, Association for Computational Linguistics, Santa Cruz, California, USA, June 1996, Online: https://www.aclweb.org/anthology/P96-1041.

Ethan A. Chi, John Hewitt, and Christopher D. Manning: 'Finding Universal Grammatical Relations in Multilingual BERT', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5564–5577, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.acl-main.493.

Nancy Chinchor, Lynette Hirschman, and David D. Lewis: 'Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)', *Computational Linguistics* 19 (3): 409–450, 1993, Online: https://www.aclweb.org/anthology/J93-3001.

Massimiliano Ciaramita and Mark Johnson: 'Supersense Tagging of Unknown Nouns in WordNet', in: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pp. 168–175, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, Online: https://doi.org/10.3115/1119355.1119377.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld: 'SPECTER: Document-level Representation Learning using Citation-informed Transformers', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.acl-main.207.

Ronan Collobert and Jason Weston: 'A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning', in: *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 160–167, ACM, New York, NY, USA, 2008, Online: http://doi.acm.org/10.1145/1390156.1390177.

Simone Conia and Roberto Navigli: 'Bridging the Gap in Multilingual Semantic Role Labeling: a Language-Agnostic Approach', in: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1396–1410, International Committee on Computational Linguistics, Barcelona, Spain (Online), December 2020, Online: https://www.aclweb.org/anthology/2020.coling-main.120.

Vincent Conitzer, Andrew Davenport, and Jayant Kalagnanam: 'Improved Bounds for Computing Kemeny Rankings', in: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pp. 620–

626, AAAI Press, 2006, Online: http://dl.acm.org/citation.cfm?id=1597538.1597638.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov: 'Unsupervised Cross-lingual Representation Learning at Scale', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Association for Computational Linguistics, Online, July 2020a, Online: https://www.aclweb.org/anthology/2020.acl-main.747.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni: 'What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Association for Computational Linguistics, Melbourne, Australia, July 2018a, Online: https://www.aclweb.org/anthology/P18-1198.

Alexis Conneau and Guillaume Lample: 'Cross-lingual Language Model Pretraining', in H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.): *Advances in Neural Information Processing Systems*, Vol. 32, pp. 7059–7069, Curran Associates, Inc., 2019, Online: https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov: 'XNLI: Evaluating Cross-lingual Sentence Representations', in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Association for Computational Linguistics, Brussels, Belgium, October-November 2018b, Online: https://www.aclweb.org/anthology/D18-1269.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov: 'Emerging Cross-lingual Structure in Pretrained Language Models', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Association for Computational Linguistics, Online, July 2020b, Online: https://www.aclweb.org/anthology/2020.acl-main.536.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan: 'GATE: an Architecture for Development of Robust HLT applications', in: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 168–175, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, July 2002, Online: https://www.aclweb.org/anthology/P02-1022.

Angel Daza and Anette Frank: 'X-SRL: A Parallel Cross-Lingual Semantic Role Labeling Dataset', in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3904–3914, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.321.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova: 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019, Online: https://www.aclweb.org/anthology/N19-1423.

Andrea Di Fabio, Simone Conia, and Roberto Navigli: 'VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling', in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 627–637, Association for Computational Linguistics, Hong Kong, China, November 2019, Online: https://www.aclweb.org/anthology/D19-1058.

William B. Dolan and Chris Brockett: 'Automatically Constructing a Corpus of Sentential Paraphrases', in: *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pp. 9–16, Asian Federation of Natural Language Processing, 2005.

David Dowty: 'Thematic Proto-Roles and Argument Selection', *Language* 76 (3): 474–496, 1991.

Timothy Dozat and Christopher D. Manning: 'Deep Biaffine Attention for Neural Dependency Parsing', *arXiv:1611.01734* 2016, Online: http://arxiv.org/abs/1611.01734.

Philipp Dufter and Hinrich Schütze: 'Identifying Elements Essential for BERT's Multilinguality', in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4423–4437, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.358.

Sebastian Ebert, Thomas Müller, and Hinrich Schütze: 'LAMB: A Good Shepherd of Morphologically Rich Languages', in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, USA, November 2016.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou: 'HotFlip: White-Box Adversarial Examples for Text Classification', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36, Association for Computational Linguistics, Melbourne, Australia, July 2018, Online: https://www.aclweb.org/anthology/P18-2006.

Steffen Eger, Erik-Lân Do Dinh, Ilia Kuznetsov, Masoud Kiaeeha, and Iryna Gurevych: 'EELECTION at SemEval-2017 Task 10: Ensemble of nEural Learners for kEyphrase ClassificaTION', in: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 942–946, Association for Computational Linguistics, Vancouver, Canada, August 2017, Online: https://www.aclweb.org/anthology/S17-2163.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg: 'Amnesic Probing: Behavioral Explanation With Amnesic Counterfactuals', *Transactions of the Association for Computational Linguistics* 9 (0): 160–175, 2021, Online: https://transacl.org/ojs/index.php/tacl/article/view/2423.

Allyson Ettinger: 'What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models', *Transactions of the Association for Computational Linguistics* 8: 34–48, 2020, Online: https://www.aclweb.org/anthology/2020.tacl-1.3.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith: 'Retrofitting Word Vectors to Semantic Lexicons', in: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1606–1615, Association for Computational Linguistics, Denver, Colorado, May–June 2015, Online: https://www.aclweb.org/anthology/N15-1184.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer: 'Problems With Evaluation of Word Embeddings Using Word Similarity Tasks', in: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 30–35, Association for Computational Linguistics, Berlin, Germany, August 2016, Online: https://www.aclweb.org/anthology/W16-2506.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber: 'Pathologies of Neural Models Make Interpretations Difficult', in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3719–3728, Association for Computational Linguistics, Brussels, Belgium, October-November 2018, Online: https://www.aclweb.org/anthology/D18-1407.

Charles J. Fillmore: 'The Case for Case', in Emmon Bach and Robert T. Harms (Eds.): *Universals in Linguistic Theory*, pp. 1–88, Holt, Rinehart and Winston, New York, 1968.

Jenny Rose Finkel, Alex Kleeman, and Christopher D. Manning: 'Efficient, Feature-based, Conditional Random Field Parsing', in: *Proceedings of ACL-08: HLT*, pp. 959–967, Association for Computational Linguistics, Columbus, Ohio, June 2008, Online: https://www.aclweb.org/anthology/P08-1109.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin: 'Placing search in context: The concept revisited', in: *Proceedings of the 10th international conference on World Wide Web*, pp. 406–414, ACM, 2001.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer: 'Large-Scale QA-SRL Parsing', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2051–2060, Association for Computational Linguistics, Melbourne, Australia, July 2018, Online: https://www.aclweb.org/anthology/P18-1191.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das: 'Semantic Role Labeling with Neural Network Factors', in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 960–970, Association for Computational Linguistics, Lisbon, Portugal, September 2015, Online: https://www.aclweb.org/anthology/D15-1112.

Lucie Flekova and Iryna Gurevych: 'Supersense Embeddings: A Unified Model for Supersense Interpretation, Prediction and Utilization', in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Vol. 1, pp. 2029–2041, Association for Computational Linguistics, 2016.

Evgeniy Gabrilovich and Shaul Markovitch: 'Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis', in: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pp. 1606–1611, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao: 'Does My Rebuttal Matter? Insights from a Major NLP Conference', in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1274–1290, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019, Online: https://www.aclweb.org/anthology/N19-1129.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer: 'AllenNLP: A Deep Semantic Natural Language Processing Platform', in: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pp. 1–6, Association for Computational Linguistics, Melbourne, Australia, July 2018, Online: https://www.aclweb.org/anthology/W18-2501.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann: 'Shortcut Learning in Deep Neural Networks', *arXiv:2004.07780* 2020.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier: 'SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD', in: *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pp. 66–74, Association for Computational Linguistics, Brussels, Belgium, November 2018, Online: https://www.aclweb.org/anthology/W18-6008.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen: 'SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity', in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2173–2182, ACL, 2016, Online: http://www.aclweb.org/anthology/D16-1235.

Daniel Gildea and Daniel Jurafsky: 'Automatic Labeling of Semantic Roles', in: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 512–520, Association for Computational Linguistics, Hong Kong, October 2000, Online: https://www.aclweb.org/anthology/P00-1065.

Ana-Maria Giuglea and Alessandro Moschitti: 'Semantic Role Labeling via FrameNet, VerbNet and PropBank', in: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 929–936, Association for Computational Linguistics, Sydney, Australia, July 2006, Online: https://www.aclweb.org/anthology/P06-1117.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka: 'Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't.', in: *Proceedings of the NAACL Student Research Workshop*, pp. 8–15, Association for Computational Linguistics, San Diego, California, June 2016, Online: https://www.aclweb.org/anthology/N16-2002.

Yoav Goldberg: 'Assessing BERT's Syntactic Abilities', *arXiv:1901.05287* 2019.

Jeffrey S. Gruber: *Studies in Lexical Relations*, Ph.D. thesis, MIT, Cambridge, MA, 1965.

James Gung and Martha Palmer: 'Predicate Representations and Polysemy in VerbNet Semantic Parsing', in: *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pp. 51–62, Association for Computational Linguistics, Groningen, The Netherlands (online), June 2021, Online: https://www.aclweb.org/anthology/2021.iwcs-1.6.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth: 'UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF', in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational*

Linguistics, pp. 580–590, Association for Computational Linguistics, Avignon, France, April 2012, Online: https://www.aclweb.org/anthology/E12-1059.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith: 'Annotation Artifacts in Natural Language Inference Data', in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, Association for Computational Linguistics, New Orleans, Louisiana, June 2018, Online: https://www.aclweb.org/anthology/N18-2017.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang: 'The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages', in: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pp. 1–18, Association for Computational Linguistics, Boulder, Colorado, June 2009, Online: https://www.aclweb.org/anthology/W09-1201.

Zellig Harris: 'Distributional structure', *Word* 10 (23): 146–162, 1954.

Tzvika Hartman, Michael D. Howell, Jeff Dean, Shlomo Hoory, Ronit Slyper, Itay Laish, Oren Gilon, Danny Vainstein, Greg S Corrado, Katherine Chou, Ming Jack Po, Jutta Williams, Scott Ellis, Gavin Bee, Avinatan Hassidim, Rony Amira, Genady Beryozkin, Idan Szpektor, and Yossi Matias: 'Customization scenarios for de-identification of clinical notes', *BMC Medical Informatics and Decision Making* 20: 2–9, 2020.

Silvana Hartmann, Iryna Gurevych, Ilia Kuznetsov, and Teresa Martin: 'Out-of-domain FrameNet Semantic Role Labeling', in: *EACL (1)*, pp. 471–482, Association for Computational Linguistics, 2017a.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych: 'Out-of-domain FrameNet Semantic Role Labeling', in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 471–482, Association for Computational Linguistics, Valencia, Spain, April 2017b, Online: https://www.aclweb.org/anthology/E17-1045.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych: 'Out-of-domain FrameNet Semantic Role Labeling', in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 471–482, Association for Computational Linguistics, Valencia, Spain, April 2017c, Online: https://www.aclweb.org/anthology/E17-1045.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer: 'Deep Semantic Role Labeling: What Works and What's Next', in: *Proceedings of the 55th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 473–483, Association for Computational Linguistics, 2017, Online: http://www.aclweb.org/anthology/P17-1044.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai: 'Syntax for Semantic Role Labeling, To Be, Or Not To Be', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2061–2071, Association for Computational Linguistics, Melbourne, Australia, July 2018, Online: https://www.aclweb.org/anthology/P18-1192.

Benjamin Heinzerling and Michael Strube: 'BPEmb: Tokenization-free Pretrained Subword Embeddings in 275 Languages', in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2989–2993, European Language Resources Association (ELRA), Miyazaki, Japan, May 2018, Online: https://www.aclweb.org/anthology/L18-1473.

John Hewitt and Percy Liang: 'Designing and Interpreting Probes with Control Tasks', in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Association for Computational Linguistics, Hong Kong, China, November 2019, Online: https://www.aclweb.org/anthology/D19-1275.

John Hewitt and Christopher D. Manning: 'A Structural Probe for Finding Syntax in Word Representations', in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019, Online: https://www.aclweb.org/anthology/N19-1419.

Felix Hill, Roi Reichart, and Anna Korhonen: 'Simlex-999: Evaluating Semantic Models with Genuine Similarity Estimation', *Computational Linguistics* 41 (4): 665–695, December 2015.

Sepp Hochreiter and Jürgen Schmidhuber: 'Long short-term memory', *Neural computation* 9 (8): 1735–1780, 1997.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly: 'Parameter-Efficient Transfer Learning for NLP', in Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.): *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 97, pp. 2790–2799, PMLR, Long Beach, California, USA, 09–15 Jun 2019, Online: http://proceedings.mlr.press/v97/houlsby19a.html.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel: ‘OntoNotes: The 90% Solution’, in: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pp. 57–60, Association for Computational Linguistics, New York City, USA, June 2006, Online: https://www.aclweb.org/anthology/N06-2015.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli: ‘SensEmbed: Learning Sense Embeddings for Word and Relational Similarity’, in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015*, pp. 95–105, Association for Computational Linguistics, 2015.

Sarthak Jain and Byron C. Wallace: ‘Attention is not Explanation’, *arXiv:1902.10186* 2019.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah: ‘What Does BERT Learn about the Structure of Language?’, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Association for Computational Linguistics, Florence, Italy, July 2019, Online: https://www.aclweb.org/anthology/P19-1356.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer: ‘Extending VerbNet with Novel Verb Classes’, in: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, pp. 1027–1032, European Language Resources Association (ELRA), Genoa, Italy, May 2006, Online: http://www.lrec-conf.org/proceedings/lrec2006/pdf/468_pdf.pdf.

Betty Kirkpatrick: *Roget’s Thesaurus of English Words and Phrases*, Penguin Reference, London, 2000.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky: ‘Revealing the Dark Secrets of BERT’, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4365–4374, Association for Computational Linguistics, Hong Kong, China, November 2019, Online: https://www.aclweb.org/anthology/D19-1445.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini: ‘Representational similarity analysis - connecting the branches of systems neuroscience’, *Frontiers in Systems Neuroscience* 2: 4, 2008, Online: https://www.frontiersin.org/article/10.3389/neuro.06.004.2008.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre: ‘Do Neural Language Models Show Preferences for Syntactic Formalisms?’, in:

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4077–4091, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.acl-main.375.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova: 'Iterative Edit-Based Unsupervised Sentence Simplification', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7918–7928, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.acl-main.707.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov: 'Measuring Bias in Contextualized Word Representations', in: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 166–172, Association for Computational Linguistics, Florence, Italy, August 2019, Online: https://www.aclweb.org/anthology/W19-3823.

Ilia Kuznetsov and Iryna Gurevych: 'Corpus-Driven Thematic Hierarchy Induction', in: *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 54–64, Association for Computational Linguistics, Brussels, Belgium, October 2018b, Online: https://www.aclweb.org/anthology/K18-1006.

Ilia Kuznetsov and Iryna Gurevych: 'From Text to Lexicon: Bridging the Gap between Word Embeddings and Lexical Resources', in: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 233–244, Association for Computational Linguistics, Santa Fe, New Mexico, USA, August 2018a, Online: https://www.aclweb.org/anthology/C18-1020.

Ilia Kuznetsov and Iryna Gurevych: 'A matter of framing: The impact of linguistic formalism on probing results', in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 171–182, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.13.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira: 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data', in: *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut: 'ALBERT: A Lite BERT for Self-supervised Learning of Language Representations', *arXiv:1909.11942* 2019.

T.K. Landauer, P.W. Foltz, and D. Laham: 'An introduction to latent semantic analysis', *Discourse processes* 25: 259–284, 1998.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang: 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics* Sep 2019, Online: http://dx.doi.org/10.1093/bioinformatics/btz682.

Beth Levin: *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, 1993, Online: https://books.google.de/books?id=6wIZWOrcBf8C.

Beth Levin: 'On Dowty's 'Thematic Proto-roles and Argument Selection'', 2019, Online: http://web.stanford.edu/~bclevin/dowty19fin.pdf.

Beth Levin and Malka Rappaport Hovav: *Argument Realization*, Research Surveys in Linguistics, Cambridge University Press, 2005.

Omer Levy and Yoav Goldberg: 'Dependency-Based Word Embeddings', in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 302–308, Association for Computational Linguistics, Baltimore, Maryland, June 2014c, Online: http://www.aclweb.org/anthology/P14-2050.

Omer Levy and Yoav Goldberg: 'Linguistic Regularities in Sparse and Explicit Word Representations', in: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 171–180, Association for Computational Linguistics, Ann Arbor, Michigan, June 2014b, Online: https://www.aclweb.org/anthology/W14-1618.

Omer Levy and Yoav Goldberg: 'Neural word embedding as implicit matrix factorization', in: *Advances in neural information processing systems*, pp. 2177–2185, 2014a.

Omer Levy, Yoav Goldberg, and Ido Dagan: 'Improving Distributional Similarity with Lessons Learned from Word Embeddings', *Transactions of the Association for Computational Linguistics* 3: 211–225, 2015, Online: https://www.aclweb.org/anthology/Q15-1016.

Dekang Lin: 'Automatic Retrieval and Clustering of Similar Words', in: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pp. 768–774, Association for Computational Linguistics, Montreal, Quebec, Canada, August 1998, Online: https://www.aclweb.org/anthology/P98-2127.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov: 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', *arXiv:1907.11692* 2019.

Edward Loper and Steven Bird: 'NLTK: The Natural Language Toolkit', in: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics -*

Volume 1, ETMTNLP 02, p. 63–70, Association for Computational Linguistics, USA, 2002, Online: https://doi.org/10.3115/1118108.1118117.

Edward Loper, Szu ting Yi, and Martha Palmer: ‘Combining lexical resources: Mapping between propbank and verbnet’, in: *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg, the Netherlands, 2007.

Thang Luong, Richard Socher, and Christopher Manning: ‘Better Word Representations with Recursive Neural Networks for Morphology’, in: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113, Association for Computational Linguistics, Sofia, Bulgaria, August 2013, Online: https://www.aclweb.org/anthology/W13-3512.

Wentao Ma, Yiming Cui, Ting Liu, Dong Wang, Shijin Wang, and Guo ping Hu: ‘Conversational Word Embedding for Retrieval-Based Dialog System’, *arXiv:2004.13249* 2020.

Xuezhe Ma and Eduard Hovy: ‘End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF’, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1064–1074, Association for Computational Linguistics, Berlin, Germany, August 2016, Online: https://www.aclweb.org/anthology/P16-1101.

Laurens van der Maaten and Geoffrey Hinton: ‘Visualizing Data using t-SNE’, *Journal of Machine Learning Research* 9: 2579–2605, 2008, Online: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky: ‘The Stanford CoreNLP natural language processing toolkit’, in: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, Association for Computational Linguistics, 2014.

Diego Marcheggiani and Ivan Titov: ‘Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling’, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1506–1515, Association for Computational Linguistics, Copenhagen, Denmark, September 2017, Online: https://www.aclweb.org/anthology/D17-1159.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz: ‘Building a Large Annotated Corpus of English: The Penn Treebank’, *Computational Linguistics* 19 (2): 313–330, 1993, Online: https://www.aclweb.org/anthology/J93-2004.

Marie-Catherine de Marneffe and Christopher D. Manning: ‘The Stanford Typed Dependencies Representation’, in: *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8, Coling 2008 Organizing Committee, Manchester, UK, August 2008, Online: https://www.aclweb.org/anthology/W08-1301.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot: 'CamemBERT: a Tasty French Language Model', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.acl-main.645.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu: 'RoBERT – A Romanian BERT Model', in: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6626–6637, International Committee on Computational Linguistics, Barcelona, Spain (Online), December 2020, Online: https://www.aclweb.org/anthology/2020.coling-main.581.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni: 'Open Language Learning for Information Extraction', in: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523–534, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher: 'Learned in Translation: Contextualized Word Vectors', in I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.): *Advances in Neural Information Processing Systems*, Vol. 30, pp. 1–12, Curran Associates, Inc., 2017, Online: https://proceedings.neurips.cc/paper/2017/file/20c86a628232a67e7bd46f76fba7ce12-Paper.pdf.

Oren Melamud, Jacob Goldberger, and Ido Dagan: 'context2vec: Learning Generic Context Embedding with Bidirectional LSTM', in: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 51–61, Association for Computational Linguistics, Berlin, Germany, August 2016, Online: https://www.aclweb.org/anthology/K16-1006.

Paola Merlo and Lonneke van der Plas: 'Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both?', in: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 288–296, Association for Computational Linguistics, 2009, Online: http://www.aclweb.org/anthology/P09-1033.

Paul Michel, Omer Levy, and Graham Neubig: 'Are Sixteen Heads Really Better than One?', in H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.): *Advances in Neural Information Processing Systems*, Vol. 32, pp. 14014–14024, Curran Associates, Inc., 2019, Online: https://proceedings.neurips.cc/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf.

Tomas Mikolov, Kai Chen, G. S. Corrado, and J. Dean: 'Efficient Estimation of Word Representations in Vector Space', *arXiv:1301.3781* 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean: 'Distributed Representations of Words and Phrases and Their Compositionality', in: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, p. 3111–3119, Curran Associates Inc., Red Hook, NY, USA, 2013b.

George A Miller: 'WordNet: a lexical database for English', *Communications of the ACM* 38 (11): 39–41, 1995.

Stuart Mitchell, Michael OSullivan, and Iain Dunning: 'PuLP: a linear programming toolkit for python', 2011, Online: http://www.optimization-online.org/DB_FILE/2011/09/3178.pdf.

Frederic Morin and Yoshua Bengio: 'Hierarchical Probabilistic Neural Network Language Model', in Robert G. Cowell and Zoubin Ghahramani (Eds.): *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. R5, pp. 246–252, PMLR, 06–08 Jan 2005, Online: http://proceedings.mlr.press/r5/morin05a.html. Reissued by PMLR on 30 March 2021.

Éva Mújdricza-Maydt, Silvana Hartmann, Iryna Gurevych, and Anette Frank: 'Combining Semantic Annotation of Word Sense & Semantic Roles: A Novel Annotation Scheme for VerbNet Roles on German Language Data', in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 3031–3038, May 2016.

Skatje Myers and Martha Palmer: 'Tuning Deep Active Learning for Semantic Role Labeling', in: *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pp. 212–221, Association for Computational Linguistics, Groningen, The Netherlands (online), June 2021, Online: https://www.aclweb.org/anthology/2021.iwcs-1.20.

Roberto Navigli and Federico Martelli: 'An overview of word and sense similarity', *Natural Language Engineering* 25 (6): 693–714, 2019.

Roberto Navigli and Simone Paolo Ponzetto: 'BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network', *Artificial Intelligence* 193: 217–250, 2012.

Allen Nie, Erin Bennett, and Noah Goodman: 'DisSent: Learning Sentence Representations from Explicit Discourse Relations', in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4497–4510, Association for Computational Linguistics, Florence, Italy, July 2019, Online: https://www.aclweb.org/anthology/P19-1442.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela: 'Adversarial NLI: A New Benchmark for Natural Language Understanding', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.acl-main.441.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman: 'Universal Dependencies v1: A Multilingual Treebank Collection', in Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1659–1666, European Language Resources Association (ELRA), Paris, France, may 2016.

Charlene Jennifer Ong, Agni Orfanoudaki, Rebecca Zhang, Francois Pierre M Caprasse, Meghan Hutch, Liang Ma, Darian Fard, Oluwafemi Balogun, Matthew I Miller, Margaret A. Minnig, Hanife Saglam, Brenton Prescott, David M. Greer, Stelios M. Smirnakis, and Dimitris Bertsimas: 'Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports', *PLoS ONE* 15, 2020.

Juri Opitz and Anette Frank: 'An Argument-Marker Model for Syntax-Agnostic Proto-Role Labeling', in: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pp. 224–234, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019, Online: https://www.aclweb.org/anthology/S19-1025.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum: 'Making fine-grained and coarse-grained sense distinctions, both manually and automatically', *Natural Language Engineering* 13 (2): 137–163, 2007.

Martha Palmer, Daniel Gildea, and Paul Kingsbury: 'The Proposition Bank: An Annotated Corpus of Semantic Roles', *Computational Linguistics* 31 (1): 71–106, 2005a, Online: https://www.aclweb.org/anthology/J05-1004.

Martha Palmer, Daniel Gildea, and Paul Kingsbury: 'The Proposition Bank: An Annotated Corpus of Semantic Roles', *Computational Linguistics* 31 (1): 71–106, March 2005b, Online: http://dx.doi.org/10.1162/0891201053630264.

Jeffrey Pennington, Richard Socher, and Christopher Manning: 'GloVe: Global vectors for word representation', in: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, Association for Computational Linguistics, Doha, Qatar, 2014.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer: 'Deep Contextualized Word Representations', in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, Association for Computational Linguistics, New Orleans, Louisiana, June 2018, Online: https://www.aclweb.org/anthology/N18-1202.

Daniel Peterson, Martha Palmer, and Shumin Wu: ‘Focusing Annotation for Semantic Role Labeling’, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pp. 4467–4471, May 2014.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller: ‘Language Models as Knowledge Bases?’, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Association for Computational Linguistics, Hong Kong, China, November 2019, Online: https://www.aclweb.org/anthology/D19-1250.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych: ‘AdapterHub: A Framework for Adapting Transformers’, *arXiv:2007.07779* 2020.

Mohammad Taher Pilehvar and Jose Camacho-Collados: ‘WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations’, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1267–1273, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019, Online: https://www.aclweb.org/anthology/N19-1128.

Telmo Pires, Eva Schlinger, and Dan Garrette: ‘How Multilingual is Multilingual BERT?’, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Association for Computational Linguistics, Florence, Italy, July 2019, Online: https://www.aclweb.org/anthology/P19-1493.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme: ‘Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation’, in: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 337–340, Association for Computational Linguistics, Brussels, Belgium, November 2018, Online: https://www.aclweb.org/anthology/W18-5441.

Sameer S. Pradhan and Nianwen Xue: ‘OntoNotes: The 90% Solution’, in: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pp. 11–12, Association for Computational Linguistics, Boulder, Colorado, May 2009, Online: https://www.aclweb.org/anthology/N09-4006.

Sai Prasanna, Anna Rogers, and Anna Rumshisky: ‘When BERT Plays the Lottery, All Tickets Are Winning’, in: *Proceedings of the 2020 Conference* *on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3208–3229, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.259.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak: 'Semantic Role Labeling Via Integer Linear Programming Inference', in: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1346–1352, COLING, Geneva, Switzerland, August 2004, Online: https://www.aclweb.org/anthology/C04-1197.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever: 'Language Models are Unsupervised Multitask Learners', 2019, Online: https://github.com/openai/gpt-2.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu: 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer', *arXiv:1910.10683* 2019.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov: 'Enhancing Unsupervised Sentence Similarity Methods with Deep Contextualised Word Representations', in: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 994–1003, INCOMA Ltd., Varna, Bulgaria, September 2019, Online: https://www.aclweb.org/anthology/R19-1115.

M. Rappaport Hovav and B. Levin: 'Deconstructing Thematic Hierarchies', in T.H. King J. Grimshaw J. Maling A. Zaenen, J. Simpson and C. Manning (Eds.): *Architectures, Rules, and Preferences: Variations on Themes*, pp. 385–402, CSLI Publications, Stanford, CA, 2007.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim: 'Visualizing and Measuring the Geometry of BERT', in H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Eds.): *Advances in Neural Information Processing Systems 32*, pp. 8594–8603, Curran Associates, Inc., 2019, Online: http://papers.nips.cc/paper/9065-visualizing-and-measuring-the-geometry-of-bert.pdf.

Nils Reimers and Iryna Gurevych: 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks', in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Association for Computational Linguistics, Hong Kong, China, November 2019, Online: https://www.aclweb.org/anthology/D19-1410.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme: 'Semantic Proto-Roles', *Transactions of the Association for Computational Linguistics* 3: 475–488, 2015a, Online: http://www.aclweb.org/anthology/Q15-1034.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme: 'Semantic Proto-Roles', *Transactions of the Association for Computational Linguistics* 3: 475–488, 2015b, Online: https://www.aclweb.org/anthology/Q15-1034.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh: 'Beyond Accuracy: Behavioral Testing of NLP Models with CheckList', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Association for Computational Linguistics, Online, July 2020, Online: https://www.aclweb.org/anthology/2020.acl-main.442.

Anna Rogers, Aleksandr Drozd, and Bofang Li: 'The (too Many) Problems of Analogical Reasoning with Word Vectors', in: *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pp. 135–148, Association for Computational Linguistics, Vancouver, Canada, August 2017, Online: https://www.aclweb.org/anthology/S17-1017.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky: 'A Primer in BERTology: What we know about how BERT works', *arXiv:2002.12327* 2020.

Michael Roth and Kristian Woodsend: 'Composition of Word Representations Improves Semantic Role Labelling', in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 407–413, Association for Computational Linguistics, 2014, Online: http://www.aclweb.org/anthology/D14-1045.

Herbert Rubenstein and John B. Goodenough: 'Contextual correlates of synonymy', *Communications of the ACM* 8 (10): 627–633, 1965, Online: http://dblp.uni-trier.de/db/journals/cacm/cacm8.html#RubensteinG65.

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme: 'Neural-Davidsonian Semantic Proto-role Labeling', in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 944–955, Association for Computational Linguistics, Brussels, Belgium, October-November 2018, Online: https://www.aclweb.org/anthology/D18-1114.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych: 'LINSPECTOR: Multilingual Probing Tasks for Word Representations', *Computational Linguistics* 46 (2): 335–385, June 2020, Online: https://www.aclweb.org/anthology/2020.cl-2.4.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi: 'Winogrande: An adversarial winograd schema challenge at scale', in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 8732–8740, 2020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf: 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter', *arXiv:1910.01108* 2020.

Karin Kipper Schuler: *VerbNet: A Broad-coverage, Comprehensive Verb Lexicon*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA, 2005.

Karin Kipper Schuler, Anna Korhonen, Neville Ryant, and Martha Palmer: 'A large-scale classification of English verbs', *Language Resources and Evaluation* 42: 21–40, 2008.

Sebastian Schuster and Christopher D. Manning: 'Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks', in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2371–2378, European Language Resources Association (ELRA), Portorož, Slovenia, May 2016, Online: https://www.aclweb.org/anthology/L16-1376.

Rico Sennrich, Barry Haddow, and Alexandra Birch: 'Neural Machine Translation of Rare Words with Subword Units', in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Association for Computational Linguistics, Berlin, Germany, August 2016, Online: https://www.aclweb.org/anthology/P16-1162.

Burr Settles: 'Biomedical named entity recognition using conditional random fields and rich feature sets', in: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA '04*, p. 104, Association for Computational Linguistics, Geneva, Switzerland, 2004, Online: http://portal.acm.org/citation.cfm?doid=1567594.1567618.

Lei Shi and Rada Mihalcea: 'Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing', in: *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'05, pp. 100–111, Springer-Verlag, Mexico City, Mexico, 2005, Online: http://dx.doi.org/10.1007/978-3-540-30586-6_9.

Peng Shi and Jimmy Lin: 'Simple BERT Models for Relation Extraction and Semantic Role Labeling', *arXiv:1904.05255* 2019.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning: 'A Gold Standard Dependency Corpus for English', in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 2897–2904, European Language Resources Association (ELRA), Reykjavik, Iceland, May 2014, Online: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1089_Paper.pdf.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman: 'Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps', *arXiv:1312.6034* 2014.

Anders Søgaard: ‘Some Languages Seem Easier to Parse Because Their Treebanks Leak’, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2765–2770, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.220.

Daniil Sorokin and Iryna Gurevych: ‘Modeling Semantics with Gated Graph Neural Networks for Knowledge Base Question Answering’, in: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3306–3317, Association for Computational Linguistics, Santa Fe, New Mexico, USA, August 2018, Online: https://www.aclweb.org/anthology/C18-1280.

Gabriel Stanovsky and Mark Hopkins: ‘Spot the Odd Man Out: Exploring the Associative Power of Lexical Resources’, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1542, Association for Computational Linguistics, Brussels, Belgium, October-November 2018, Online: https://www.aclweb.org/anthology/D18-1182.

Kevin Stowe, Jenette Preciado, Kathryn Conger, Susan Windisch Brown, Ghazaleh Kazeminejad, and Martha Palmer: ‘SemLink 2.0: Chasing Lexical Resources’, in: *30th Annual Meeting of the Association for Computational Linguistics*, pp. 222–227, Association for Computational Linguistics, Gronigen, Netherlands, 2021, Online: https://iwcs2021.github.io/proceedings/iwcs/pdf/2021.iwcs-1.21.pdf.

Emma Strubell, Ananya Ganesh, and Andrew McCallum: ‘Energy and Policy Considerations for Deep Learning in NLP’, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, Association for Computational Linguistics, Florence, Italy, July 2019, Online: https://www.aclweb.org/anthology/P19-1355.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum: ‘Linguistically-Informed Self-Attention for Semantic Role Labeling’, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5027–5038, Association for Computational Linguistics, Brussels, Belgium, October-November 2018, Online: https://www.aclweb.org/anthology/D18-1548.

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin: ‘Adaptive Attention Span in Transformers’, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 331–335, Association for Computational Linguistics, Florence, Italy, July 2019, Online: https://www.aclweb.org/anthology/P19-1032.

Weiwei Sun, Zhifang Sui, and Meng Wang: ‘Prediction of Thematic Rank for Structured Semantic Role Labeling’, in: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 253–256, Association for Computational Linguistics, 2009, Online: http://www.aclweb.org/anthology/P09-2064.

Chris Sweeney and Maryam Najafian: 'A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings', in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1662–1667, Association for Computational Linguistics, Florence, Italy, July 2019, Online: https://www.aclweb.org/anthology/P19-1162.

Yi Chern Tan and L. Elisa Celis: 'Assessing Social and Intersectional Biases in Contextualized Word Representations', in H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.): *Advances in Neural Information Processing Systems 32*, pp. 13230–13241, Curran Associates, Inc., 2019.

Ian Tenney, Dipanjan Das, and Ellie Pavlick: 'BERT Rediscovers the Classical NLP Pipeline', in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Association for Computational Linguistics, Florence, Italy, July 2019a, Online: https://www.aclweb.org/anthology/P19-1452.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick: 'What do you learn from context? Probing for sentence structure in contextualized word representations', *7th International Conference on Learning Representations, ICLR 2019* pp. 1–17, 2019b, Online: https://arxiv.org/pdf/1905.06316.pdf.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning: 'A Global Joint Model for Semantic Role Labeling', *Computational Linguistics* 34 (2): 161–191, 2008, Online: https://aclanthology.org/J08-2002.

Andrew Trask, Phil Michalak, and John Liu: 'sense2vec – A fast and accurate method for word sense disambiguation in neural word embeddings', *arXiv:1511.06388* 2015.

Peter D Turney: 'Similarity of semantic relations', *Computational Linguistics* 32 (3): 379–416, 2006.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin: 'Attention is All you Need', in I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.): *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, Curran Associates, Inc., 2017, Online: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Jesse Vig: 'A Multiscale Visualization of Attention in the Transformer Model', in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 37–42, Association for Computational Linguistics, Florence, Italy, July 2019, Online: https://www.aclweb.org/anthology/P19-3007.

Luke Vilnis and Andrew McCallum: 'Word representations via gaussian embedding', *arXiv:1412.6623* 2014.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov: 'Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned', in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Association for Computational Linguistics, Florence, Italy, July 2019, Online: https://www.aclweb.org/anthology/P19-1580.

Elena Voita and Ivan Titov: 'Information-Theoretic Probing with Minimum Description Length', in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.14.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim: 'BERTje: A Dutch BERT Model', *arXiv:1912.09582* December 2019.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim: 'What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models', in: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4339–4350, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.findings-emnlp.389.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen: 'Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity', *Computational Linguistics* 46 (4): 847–897, 02 2020, Online: https://doi.org/10.1162/coli_a_00391.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen: 'Cross-Lingual Induction and Transfer of Verb Classes Based on Word Vector Space Specialisation', in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2546–2558, Association for Computational Linguistics, Copenhagen, Denmark, September 2017a, Online: https://www.aclweb.org/anthology/D17-1270.

Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen: 'Morph-fitting: Fine-Tuning Word Vector Spaces with Simple Language-Specific Rules', in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 56–68, Association for Computational Linguistics, Vancouver, Canada, July 2017b, Online: http://aclweb.org/anthology/P17-1006.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen: 'Probing Pretrained Language Models for Lexical Semantics', in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7222–7240, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.586.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh: 'AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models', in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 7–12, Association for Computational Linguistics, Hong Kong, China, November 2019, Online: https://www.aclweb.org/anthology/D19-3002.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman: 'SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems', *CoRR* abs/1905.00537, 2019, Online: http://arxiv.org/abs/1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman: 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding', in: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Association for Computational Linguistics, Brussels, Belgium, November 2018, Online: https://www.aclweb.org/anthology/W18-5446.

Le Wang, Xiao kang Wang, Juan juan Peng, and Jian qiang Wang: 'The differences in hotel selection among various types of travellers: A comparative analysis with a useful bounded rationality behavioural decision support model', *Tourism Management* 76: 103961, 2020, Online: http://www.sciencedirect.com/science/article/pii/S0261517719301591.

Alex Warstadt and Samuel R. Bowman: 'Linguistic Analysis of Pretrained Sentence Encoders with Acceptability Judgments', *arXiv:1901.03438* 2020.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman: 'Neural Network Acceptability Judgments', *arXiv:1805.12471* 2019.

Aaron Steven White, Kyle Rawlins, and Benjamin Van Durme: 'The Semantic Proto-Role Linking Model', in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 2, pp. 92–98, Association for Computational Linguistics, 2017, Online: http://aclweb.org/anthology/E/E17/E17-2015.pdf.

Aaron Steven White, Drew Reisinger, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme: 'Computational linking theory', *arXiv:1610.02544* 2016a.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme: 'Universal Decompositional Semantics on Universal Dependencies', in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1713–1723, Association for Computational Linguistics, Austin, Texas, November 2016b, Online: https://www.aclweb.org/anthology/D16-1177.

Aaron Steven White, Elias Stengel-Eskin, Siddharth Vashishtha, Venkata Subrahmanyan Govindarajan, Dee Ann Reisinger, Tim Vieira, Keisuke Sakaguchi, Sheng Zhang, Francis Ferraro, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme: 'The Universal Decompositional Semantics Dataset and Decomp Toolkit', in: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5698–5707, European Language Resources Association, Marseille, France, May 2020, Online: https://www.aclweb.org/anthology/2020.lrec-1.699.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann: 'Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings', *arXiv:1909.10430* 2019.

Sarah Wiegreffe and Yuval Pinter: 'Attention is not not Explanation', in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Association for Computational Linguistics, Hong Kong, China, November 2019, Online: https://www.aclweb.org/anthology/D19-1002.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu: 'Charagram: Embedding Words and Sentences via Character n-grams', in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1504–1515, Association for Computational Linguistics, Austin, Texas, US, November 2016, Online: https://aclweb.org/anthology/D16-1157.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew: 'HuggingFace's Transformers: State-of-the-art Natural Language Processing', *arXiv:1910.03771* 2019.

Y. Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Q. Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, Taku Kudo, H. Kazawa, K. Stevens, G. Kurian, Nishant Patil, W. Wang, C. Young, J. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, G. S. Corrado, Macduff Hughes, and J. Dean: 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation', *ArXiv* abs/1609.08144, 2016.

Patrick Xia, Shijie Wu, and Benjamin Van Durme: 'Which *BERT? A Survey Organizing Contextualized Encoders', in: *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7516–7533, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.608.

Nianwen Xue and Martha Palmer: 'Calibrating Features for Semantic Role Labeling', in: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 88–94, Association for Computational Linguistics, Barcelona, Spain, July 2004, Online: https://www.aclweb.org/anthology/W04-3212.

Yi Yang and Arzoo Katiyar: 'Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning', in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6365–6375, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.516.

Benat Zapirain, Eneko Agirre, and Lluìs Marquez: 'Robustness and Generalization of Role Sets: PropBank vs. VerbNet', in: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, number June, pp. 550–558, Association for Computational Linguistics, 2008, Online: http://www.aclweb.org/anthology/P08-1063.

Amir Zeldes: 'The GUM Corpus: Creating Multilayer Resources in the Classroom', *Language Resources and Evaluation* 51 (3): 581–612, 2017.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov: 'CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies', in: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1–21, Association for Computational Linguistics, Brussels, Belgium, October 2018, Online: https://www.aclweb.org/anthology/K18-2001.

Zining Zhu and Frank Rudzicz: 'An information theoretic view on selecting linguistic probes', in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9251–9262, Association for Computational Linguistics, Online, November 2020, Online: https://www.aclweb.org/anthology/2020.emnlp-main.744.