Eötvös Loránd University
Faculty of Arts

Theses of PhD. Dissertation

Gábor Recski

Computational methods in semantics

Doctoral School of Linguistics
Gábor Tolcsvai Nagy MHAS

Theoretical Linguistics Doctoral Programme
Zoltán Bánréti CSc.

Members of the Committee:
Ferenc Kiefer MHAS (chair)
Péter Rebrus PhD.
Veronika Vincze PhD.
Gábor Alberti DSc.
András Komlósy CSc.

Supervisor:
András Kornai DSc.

Budapest, 2016

Introduction

The dissertation presents the `4lang` software library, which builds `4lang`-style semantic representations from raw English and Hungarian text, processes entries of monolingual dictionaries to acquire definitions of any headword, and includes a module for extracting word similarity features from pairs of definition graphs. The work also presents two competitive systems for measuring semantic similarity based on representations built using the `4lang` library, one of which is the current state of the art algorithm on the popular SimLex benchmark for measuring the similarity of English word pairs. This booklet lists the theses of the dissertation, outlines the structure of the full work, enumerates external contributions to each system, provides links to all software, and finally presents a brief summary of each of the theses.

Theses

The main theses of the disseration are the following:

(T1) The `text_to_4lang` tool for building `4lang`-style semantic representations from English and Hungarian raw text

(T2) The `dict_to_4lang` tool for building `4lang` definition graphs from monolingual dictionaries of English and Hungarian

(T3) A competitive system for measuring the semantic similarity of English sentence pairs using definition graphs built by `dict_to_4lang`

(T4) The current state of the art algorithm for measuring the semantic similarity of English word pairs using features extracted from `4lang` graphs

Structure of the dissertation

Chapter 2 gives a short review of existing theories of word meaning, with special focus on their applicability to natural language processing. Chapter 3 provides an overview of the `4lang` formalism for modeling meaning, but will not attempt a full discussion, since the `4lang` formalism is the product of joint work by half a dozen researchers (Kornai et al., 2015), rather than being a contribution of this thesis. Chapter 4 presents the `dep_to_4lang` pipeline, which creates `4lang`-style meaning representations from running text, Chapter 5 describes its application to monolingual dictionary definitions, `dict_to_4lang`, used to create large concept lexica automatically. Chapter 6 presents applications of the `text_to_4lang` module to various tasks in Computational Semantics, including a competitive system for measuring semantic textual similarity (STS) (Recski & Ács, 2015), and a hybrid ML-based system for measuring the similarity of English word pairs, which at the time of submission is the top-scoring algorithm on the popular SimLex benchmark dataset (Recski et al., 2016). The chapter also briefly describes an experimental framework for natural language understanding (Nemeskey et al., 2013) based on `4lang` representations. Chapter 7 presents the architecture of the ca.

System	Code	Main publication
`4lang`	`github.com/kornai/4lang`	(Recski, 2016)
`pymachine`	`github.com/kornai/pymachine`	
`semeval`	`github.com/juditacs/semeval`	(Recski & Ács, 2015)
`4lang`	`github.com/recski/wordsim`	(Recski et al., 2016)

Table 1: Software libraries presented in this thesis

3000-line `4lang` codebase, serving both as an overview of how the main tools presented in the thesis are implemented and as comprehensive software documentation. Finally, Chapter 8 discusses our plans for future applications.

Contributions

The `4lang` principles outlined in Chapter 3 are the result of collaboration with current and former members of the Research Group for Mathematical Linguistics at the Hungarian Academy of Sciences: Judit Ács, Gábor Borbély, András Kornai, Márton Makrai, Dávid Nemeskey, Katalin Pajkossy, and Attila Zséder. The systems presented in Chapters 4 and 5 constitute the author's work with only minor exceptions: the functions performing graph expansion (Section 5.3) are a result of joint work with Gábor Borbély, and a parser for the Collins Dictionary was contributed by Attila Bolevácz. The SemEval system presented in Section 6.1 were built in collaboration with Judit Ács, the more recent `wordsim` system presented in Section 6.2 is a result of joint work with Eszter Iklódi (Department of Automation and Applied Informatics, Budapest University of Technology and Economics), key ML components were contributed by Katalin Pajkossy. The experimental systems described in Section 6.3 were implemented together with Dávid Nemeskey and Attila Zséder.

Software

All software presented in the thesis is available for download under an MIT license, URLs are listed in Table 1. The `text_to_4lang` and `dict_to_4lang` tools (T1-T2) are parts of the `4lang` library, some dependencies are included in the package `pymachine`. The state of the `4lang` codebase at the time of submission of this thesis is preserved in the branch `recski_thesis`. The sentence similarity system (T3) is preserved in the `semeval` repository, the word similarity system (T4) is part of the `wordsim` package. All external dependencies of these systems are freely downloadable under various open-source licenses.

(T1) The text_to_4lang tool for building 4lang-style semantic representations from English and Hungarian raw text

text_to_4lang is a module of the 4lang library and maps raw text to 4lang semantic representations by invoking standard dependency parsers and postprocessing their output. text_to_4lang relies on the dep_to_4lang module for mapping dependencies output by either the Stanford Parser or magyarlanc to subgraphs over 4lang concepts. Some specific structures such as coordination or copular sentences are handled by ad-hoc rules for postprocessing dependencies. Although suffering from errors made by each syntactic parser and currently limited to decoding patterns within their scope, text_to_4lang yields high-quality representations for simple sentences and individual phrases, serving as the basis for all applications described in the thesis.

(T2) The dict_to_4lang tool for building 4lang definition graphs from monolingual dictionaries of English and Hungarian

The dict_to_4lang module of the 4lang library builds 4lang-style concept definitions by processing entries of monolingual dictionaries. dict_to_4lang is an application of text_to_4lang but extends its functionalities with preprocessing steps specific to each of the five datasources currently supported (3 English and 2 Hungarian dictionaries). Manual evaluation of automatically built graphs reveal that dict_to_4lang produces highly accurate definitions for over 60% of English headwords and 80% of Hungarian headwords. Once definition graphs are available for virtually all words of English and Hungarian, 4lang representations can be *expanded*: relations from definitions of each concept can be added to a 4lang graph, and any representation can thus be traced back to an arbitrarily small set of basic concepts.

(T3) A competitive system for measuring the semantic similarity of English sentence pairs using definition graphs built by dict_to_4lang

Our system for measuring the semantic similarity of pairs of English sentences is a reimplementation of an architecture that has been used successfully at multiple Semeval competitions and that relies on multiple metrics of word similarity to derive overall scores for sentence pairs. Our system uses Support Vector Regression to combine various metrics, including features defined over pairs of 4lang definition graphs. Some 4lang-based features measure the ratio of overlapping subgraphs, others encode particular configurations that correlate with semantic similarity. Our top-scoring system achieves a mean Pearson correlation of 0.78 on the evaluation dataset of 2015 Semeval task for Textual Similarity, placing 12th among 78 submitted systems.

(T4) The current state of the art algorithm for measuring the semantic similarity of English word pairs using features extracted from 4lang graphs

The wordsim system for measuring the semantic similarity of English word pairs achieves the highest correlation with human annotators on the SimLex dataset, which has become the standard benchmark for this task since its introduction in early 2015. Our model combines various distributional models of word similarity, features based on WordNet representations, and features defined over pairs of 4lang definitions, some of which have also been used by the system described in (T3). We show that a model combining a selection of specialized word embeddings already outperforms most existing systems, but including features 4lang-based features further increases performance by a significant margin and to a larger extent than those based on WordNet. Our top correlation of 0.76 is higher than any published system that we are aware of, well beyond the average inter-annotator agreement of 0.67, and close to the 0.78 average correlation between a human rater and the average of all other ratings, showing that our system has achieved near-human performance on this benchmark.

References

Kornai, A., Ács, J., Makrai, M., Nemeskey, D. M., Pajkossy, K., & Recski, G. (2015). Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)* (pp. 165–175). Denver, Colorado: Association for Computational Linguistics.

Nemeskey, D., Recski, G., Makrai, M., Zséder, A., & Kornai, A. (2013). Spreading activation in language understanding. In *Proceedings of the 9th International Conference on Computer Science and Information Technologies (CSIT 2013)* (pp. 140–143). Yerevan, Armenia: Springer.

Recski, G. (2016). Building concept graphs from monolingual dictionary entries. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, Slovenia: European Language Resources Association (ELRA).

Recski, G., & Ács, J. (2015). MathLingBudapest: Concept networks for semantic similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 543–547). Denver, Colorado: Association for Computational Linguistics.

Recski, G., Iklódi, E., Pajkossy, K., & Kornai, A. (2016). Measuring semantic similarity of words using concept networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP* (pp. 193–200). Berlin, Germany: Association for Computational Linguistics.