Empirical Exploration of Factors Related to Tacit Knowledge Acquisition

Empirical exploration of factors related to tacit knowledge acquisition

暗黙知の獲得に関連する要因の経験的調査

Diana Galván Sosa

Graduate School of Information Sciences
Tohoku University

This dissertation is submitted for the degree of
Doctor of Information Science

January 2021

Acknowledgements

Even though pursuing a Ph.D. was something I didn't see myself doing 6 years ago when I arrived to Japan, I couldn't be happier to have continued with my graduate studies. This decision was strongly influenced by Prof. Kentaro Inui, to whom I owe everything I know about NLP. He has guided me from the day I requested to join his group and through all my graduate studies, always encouraging me to don't give up and follow what I am passion about. Thank you so much for your support and patience. I am also very thankful to my mentor Koji Matsuda for always being by my side helping push myself through situations, no matter what the problem was. He along with Prof. Jun Suzuki taught me a lot through our discussions. They were my main collaborators, but I would like to extend my gratitude to the other members of the lab; they all contributed in different ways to shape the person who I am today.

As everything I do in my life, I dedicate this work to my sisters and my parents. It has not been easy to be away from home for so long but despite the distance, your unconditional support has helped me not feel alone. I can't stop mentioning my dear friends for always being there for me. Special thanks to my *coach* Isabel and my best friend, Paola, who I reached out whenever I felt I couldn't go on. Thank you both for always being there for good or bad, helping me go through the difficult moments and celebrating with me every success as if it were yours.

The last year of my Ph.D. program was the more challenging and I wouldn't have been able to go through it if it weren't for my wonderful novia Jade. I lost count of the many times she was there to hold me and to remind me that everything was going to be ok. I have no words to thank you for being such a great friend and partner. I look forward to keep sharing my life with you.

Last, but not least, I would like to thank Japan's Ministry of the Education, Culture, Sports, Science and Technology (MEXT) for financing my graduate studies at Tohoku University.

Abstract

Artificial intelligence is evolving at an amazingly fast pace. The combination of deep neural networks and the great deal of annotated datasets and large-scale resources like Wikipedia have made possible for today's systems to handle tasks that require the use of a tool unique to humans: language. Systems are trained on these natural language data and as a result, they are able to acquire knowledge about the order (syntax) and meaning (semantics) of words. Additionally, they also learn general knowledge about the world around us. Humans use language as a tool for communicating what they see, think and feel. The information we report is varied; it ranges from the characteristics of a PERSON (e.g., *Barack Obama*, *Lionel Messi*) or a PLACE (e.g., *The United States of America*, *Barcelona*, *Camp Nou*) to our opinion about a product or our stance towards a particular topic. Text data is rich in this type of information and it has definitely had an impact on natural language processing (NLP). However, linguistic competence also requires *tacit* knowledge, a type of knowledge about things that are difficult to explain, like the notion of time or common sense.

This thesis presents an exploratory study of NLP systems capability to handle tacit knowledge. We aim to answer to what degree have they acquired such knowledge and how do different text sources like encyclopedic articles, the entries of a knowledge base and descriptions about real world images contribute to tacit knowledge acquisition. Our work addresses two tasks: temporal relation extraction, in which knowledge about the duration of an event is crucial, and commonsense-based machine reading comprehension. For each task, we chose an existing state-of-the-art system and deliver a deep analysis of its performance.

Table of contents

List of figures vii

List of tables ix

List of publications x

1 Introduction 1

1.1 Background . . . 1

1.1.1 Computers and language . . . 1

1.1.2 Tacit knowledge . . . 2

1.2 Contribution . . . 4

1.3 Thesis overview . . . 4

2 Empirical exploration of the challenges in temporal relation extraction from clinical text 5

2.1 Introduction . . . 5

2.1.1 Clinical TempEval challenges . . . 6

2.1.2 Previous work . . . 7

2.2 Methods . . . 8

2.2.1 From relation extraction to temporal relation extraction . . . 8

2.2.2 Experimental settings . . . 9

2.3 Results . . . 11

2.4 Discussions . . . 12

2.4.1 Error analysis . . . 12

2.4.2 Temporal relations and aspectual classes . . . 14

2.4.3 Temporality of nominal events . . . 18

2.4.4 Precision and recall imbalance . . . 21

2.5 Conclusions . . . 23

3 Evaluation of image descriptions for commonsense reasoning in machine reading comprehension 25
3.1 Introduction . . . 25
3.2 Related work . . . 26
3.3 Approach . . . 27
3.4 Experiments . . . 28
3.4.1 Data . . . 28
3.4.2 Models . . . 29
3.5 Results . . . 31
3.6 Conclusion . . . 32

4 Comparing the content of two text sources and their impact on commonsense machine reading comprehension 34
4.1 Introduction . . . 34
4.2 Automating external knowledge retrieval . . . 36
4.2.1 Question-answer retrieval . . . 37
4.2.2 Question-passage-answer retrieval . . . 37
4.3 Experiments . . . 38
4.3.1 Indexes . . . 38
4.3.2 Data . . . 38
4.3.3 Models . . . 39
4.4 Results . . . 40
4.5 Evaluation . . . 41
4.5.1 Retrieval inspection . . . 41
4.5.2 Model inspection . . . 42
4.6 Conclusion . . . 46

5 Conclusion 48

References 50

Appendix A Miwa and Bansal's model adaptation 56
A.1 Settings . . . 56
A.1.1 Sentence-level annotations . . . 56
A.1.2 Implementation and training . . . 57
A.2 Multi-classification performance . . . 57
A.2.1 In-domain word embeddings . . . 57

A.2.2 Down-sampling negative examples 58

List of figures

2.1 Example temporal relation annotation with and without using narrative containers. 6
2.2 Confusion matrix of our multi-class classification model with PubMed word embeddings on the dev set. 13
2.3 Vendler's four-way classification. Arrows represent an indefinite time interval, solid lines indicate a homogeneous duration, and dashed lines indicate a dynamic duration. An X is used to represent a situation's natural end point. Abbreviations: C–Clear and NC–Not Clear. 15
2.4 Allen and Vendler's interval representation of OVERLAP and CONTAINS relations. A- / B- and A+ / B+ represent the start and end of an event, respectively. Filled-dots represent clear start points while an empty-dot represent a not-clear start point. 16
2.5 CONTEXTUAL ASPECT attribute values by set 18

3.1 Visualization of MCScript2.0 original data split and our data split. 28
3.2 Example of three selected and one removed commonsense questions from two MCScript2.0 instances. 29
3.3 Two input/output examples. In the top example, region descriptions were not helpful to chose the correct answer candidate. In the bottom example, they were. 30
3.4 Retrieval process for one of the questions BERT answered incorrectly. Identifying the GOING SHOPPING scenario, querying Visual Genome and selecting the most related region descriptions to the scenario was manually done. . . 31
3.5 Examples of questions from the unanswerable set and one of the manually selected region descriptions from Visual Genome. 32

4.1 Example text fragment from MCScript2.0 35

4.2 Visualization of MCScript2.0 original data split. In the right, a visualization of how we derived the data used in Chapter 3. Below, the data split used in the current chapter. 39
4.3 Distribution of VISUAL COMMONSENSE QUESTIONS in our dev set with 740 commonsense questions total. 41
4.4 Overall impact on the 740 commonsense questions from the dev set. The impact was positive, negative or neutral depending on the change on the logit value of the CORRECT ANSWER CANDIDATE. 44
4.5 Overall impact on the 279 VISUAL COMMONSENSE QUESTIONS from the dev set. The impact was positive, negative or neutral depending on the change on the logit value of the CORRECT ANSWER CANDIDATE 45

List of tables

2.1 Label distribution of pre-processed dataset for binary classification. 9
2.2 Label distribution of pre-processed dataset for multi-class classification. . . 9
2.3 Performance of systems and humans on identifying CONTAINS relations. Our results come from five different random seeds. 10
2.4 Results of the three multi-class classification experiments and Leeuwenberg and Moens's 2017 Structured Perceptron (SP) best results on the THYME test set. The SP results were reproduced from the original paper. The results come from five different random seeds. FNE refers to filtered negative examples. 12
2.5 Sample of the analyzed misclassified sentences by our system. e_1 and e_2 are shown in bold and italics, respectively. 17
2.6 Distribution of misclassified CONTAINS and OVERLAP pairs by type of TLINK (left) and EVENT type (right). Abbreviations: E–EVENT, T–TIMEX3, V–Verb and NV–Non-Verb 19
2.7 Results of our multi-class classification experiments on the THYME test set. Our results come from five different random seeds. *Without None* refers to training without the None class. 21
2.8 Results of our multi-class classification experiments on the THYME test set. The first results are results obtained when oversampling only OVERLAP, and the subsequent results are obtained when oversampling BEFORE, BEGINS-ON, ENDS-ON, and OVERLAP. 23

3.1 Accuracy of BERT baseline and our manually visually enhanced BERT in both MCScript2.0 development and test sets. The results come from three different random seeds. 31

4.1 Distribution of MCScript2.0 instances and questions on each data split. . . 38

4.2 Accuracy on commonsense questions from MCScript2.0. The results are the average of three runs using different random seeds. 40
4.3 Sample of meaningful image descriptions and surface texts. The correct answer candidate is underlined. A dash line (-) indicates there was no meaningful entry found in the top 50 results. 43
4.4 5-fold cross-validation accuracy. After the + sign, the name of the index that was queried. 46

List of publications

Journal Paper (Refereed):

1. Diana Galvan-Sosa, Koji Matsuda, Naoaki Okazaki, Kentaro Inui. Empirical exploration of the challenges in temporal relation extraction from clinical text. Journal of Natural Language Processing, Vol.27 No. 2, June 2020.

International Conferences/Workshop Papers (Refereed):

1. Diana Galvan, Naoaki Okazaki, Koji Matsuda and Kentaro Inui. Investigating the Challenges of Temporal Relation Extraction from Clinical Text. In Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pp.55-64, October 2018.

2. Diana Galvan-Sosa, Jun Suzuki, Kyosuke Nishida, Koji Matsuda, Kentaro Inui. Seeing the world through text: Evaluating image descriptions for commonsense reasoning in machine reading comprehension. In Proceedings of the Beyond Vision and LANguage: inTEgreating Real-world kNowledge (LANTERN), pp.23-29, December 2020.

Other Publications (Not refereed):

1. Diana Galvan-Sosa, 西田京介, 松田耕史, 鈴木潤, 乾健太郎. テキストを通して世界を見る: 機械読解における常識的推論のための画像説明文の評価. 言語処理学会第 26 回年次大会, pp.593-596, March 2020.

Chapter 1

Introduction

1.1 Background

1.1.1 Computers and language

Computers have an obvious advantage over humans: they are capable of processing large amounts of information in a short time. It is thanks to computational systems that tasks like data mining and analysis have been completely automated and they can be performed on a large scale. As computers continued to progress rapidly, so did the interest in automating the processing of natural language data. Examples of Natural Language Processing (NLP) tasks include:

1. **Sentiment Analysis**: The task of identifying whether the content of a text (sentence or paragraph) is positive or negative.

2. **Machine Translation**: The task of expressing the sense of a source text in another language.

3. **Question Answering**: The task of providing an answer to a given question.

NLP systems that are able to handle the aforementioned tasks are already a reality; (1) is commonly used by e-commerce services like Amazon, Google implementation of (2) is widely used to translate websites in foreign languages like German, French or Spanish into English (and vice versa) and (3) is implemented in various search engines to allow users to query information using natural language instead of defining a set of keywords (e.g., *How can I make banana bread?* instead of *banana bread recipe*).

NLP has been growing at a fast pace thanks to deep neural networks or *deep learning* for short. Deep learning involves a network in which artificial neurons–typically thousands,

millions, or many more of them–are stacked at least several layers deep. The artificial neurons in the first layer pass information to the second, the second to the third, and so on, until the final layer outputs some values (Krohn et al., 2019). Through this process, systems are able to learn features that represent the meaning of words. However, language processing requires more than knowledge about words and their compositional rules.

Language is a tool to interact, a means to convey thoughts, ideas, concepts or even feelings. Consider the following sentence:

"Romeo and Juliet" is one of Shakespeare's early tragedies.

The above example, taken from Ovchinnikova (2012), shows that we need **world knowledge** to make sense of what the sentence means. It is necessary to rely in our general conception of the world to identify that the name *Shakespeare* refers to the famous English playwright *William Shakespeare*. Knowing this, it is easy to conclude that *tragedy*, in this context, refers to a work of art rather than to a dramatic event. Our background knowledge also helps us to understand that the time expression *early* is used to refer to an event relative to the lifetime of Shakespeare. Therefore, we conclude that Shakespeare wrote *Romeo and Juliet* when he was young.

An NLP system would be able to make the aforementioned inferences by being trained on Wikipedia articles about *Shakespeare*, *Art* or *Plays* to learn facts like *Shakespeare is a playwright* and *playwrights write plays*. However, while there are some concepts and facts about the world that can be made explicit and therefore, available for a NLP system to learn from it, there are some others that are hard to explain. Such knowledge is commonly referred to as **tacit knowledge**.

1.1.2 Tacit knowledge

The *Chambers dictionary* defines *tacit* as "unspoken", "understood but not actually stated; implied". Tacit is essentially the opposite to explicit (Collins, 2010) and is commonly used interchangeably with *implicit* despite having slight but important differences. Both *tacit* and *implicit* convey the meaning of something that is not directly expressed, but the term tacit implies that something is not being mentioned because it is difficult to find words to describe it. *Tacit* began to be used by Michael Polanyi, a philosopher who made the assertion that *"we can know more than we can tell"*. He used the term to imply that there are things that one *cannot* explain, rather than there are things that we *can* explain but is *hard* to do so. In this work, we stick to the dictionary definition of tacit which implies the latter, but we agree on Polanyi's claim that there is knowledge that we are aware of, but we are not sure how

did we get to know it. Polanyi's classic example of tacit knowledge is the ability of *riding a bicycle*. Bike-riding is tacit because it is an activity that we are able to do without being given any instructions and similarly, we can claim that we know how to ride a bicycle even if we cannot make explicit the rules of riding. We sure can come up with something like *"hold the handle and start pedaling"*, but explaining how to balance your body while riding does not come as easy.

There are two types of knowledge necessary for language understanding that we can label as being tacit: **temporal knowledge** and **commonsense knowledge**. Extending Polanyi's *riding a bicycle* example, we are not only aware of how to ride, but also how long does the action of riding usually lasts and that we need, of course, to get a bicycle and sit on top of it in order to ride it. The former has to do with our knowledge about the *duration* of events and the latter with knowledge so obvious that we assume other people to have it. In our previous example about Shakespeare, temporal knowledge is what helps us understand the meaning of *early*. Similarly, if we were to process a text about riding a bike with a the sentence *I did not enjoy the ride, it was very uncomfortable* our common sense would tell us that the sentence means that the bike's seat was not comfortable.

When it comes to temporal knowledge, there are two key concepts that frame our notion of time: *duration* and *sequence*. As shown in Chapter 2, the concept of duration is particularly difficult to represent not only for humans, but also for computers. There is no doubt, though, that a human is aware of how long does an event last, so why is it so difficult to come up with an accurate time representation? It is hard to precise how does a human learn about the duration of any action and it is even harder to fix a time interval for each one of them. This is why temporal knowledge is considered to be tacit. When we try to think about other types of knowledge as inherent to us as our notion of time, it is impossible not to think of *commonsense*. Commonsense knowledge is a broader term that besides temporal knowledge, it includes, among others, physical (i.e., shape and color of objects), spatial (i.e., location of objects) and social knowledge. All of them are characterized by a set of fundamental assumptions and expectations regarding the nature of the world (Kulyk, 2006). Commonsense knowledge is so hard to explain that there is not a unique definition of it, let alone a clear understanding of how do we come to develop it. Just like temporal knowledge, commonsense knowledge is tacit.

It is clear that knowledge is critical to achieve true human-level language understanding. We know that current NLP systems do have knowledge, to some degree, given that they are able to perform language-related tasks. However, we cannot say that language processing is solved. There still are machine translations that look unnatural or simple questions that

a system cannot seem to answer probably because they are lacking *tacit knowledge*. This motivated us to conduct the present exploratory study.

1.2 Contribution

The contribution of this thesis is roughly divided into the following points:

1. We explore to what degree do state-of-the-art NLP systems handle tacit knowledge. To this end, we evaluate the performance of two systems on two tacit knowledge-sensitive tasks (temporal relation extraction and commonsense machine reading comprehension) and deliver a deep analysis of the results.

2. In addition to our analysis of a system's performance, we explore how two different text sources contribute to learn tacit knowledge. One of the sources is the text generated to describe an image, which has not been tested in conjunction with a system on a downstream task. The second are entries from a well-known commonsense knowledge base. We extrinsically evaluate their content, designing a retrieval module that extracts relevant information from either source and incorporates it to a machine reading comprehension system.

3. By applying the method proposed in above, we identified what makes the temporal relation extraction task so challenging to NLP systems. We were also able to measure how different text sources contribute to commonsense knowledge acquisition.

1.3 Thesis overview

In this section, we explain the structure of this thesis. In Chapter 2, we target the task of temporal relation extraction, a task that the NLP community recognizes as one of the most difficult ones. This chapter is dedicated to find the reason(s) why this task is so challenging. In Chapter 3 we focus on commonsense knowledge about every day activities. We explore to what degree descriptions of real world images help a system improve its performance on a machine reading comprehension task. In Chapter 4, we further explore the content of image descriptions, comparing it against a commonsense knowledge base. Finally, in Chapter 5, we review the summary of the above research and its contribution.

Chapter 2

Empirical exploration of the challenges in temporal relation extraction from clinical text

In this chapter, we focus on temporal knowledge. We present several experiments on one of the most challenging NLP tasks to get a better understanding of what is it about temporal knowledge that a system finds difficult to process and how could we alleviate such a problem.

2.1 Introduction

Human reasoning has to do with time. High-level cognition concepts, such as *duration* and *sequence*, influence the structure of human interaction with the external world. Temporal reasoning is a fundamental ability not only in humans but also in intelligent systems. In Natural Language Processing (NLP), Temporal Information Extraction (TIE) is an active research area where the ultimate goal is to be able to represent the development of a story over time. This is key to text processing tasks including question answering (UzZaman et al., 2012) and text summarization (Jung et al., 2011), and it follows the traditional pipeline of named entity recognition and relation extraction separately. In a temporal context, entities are typically classified as either events or time expressions and temporal relations describe how their time intervals interact, assuming a linear model of time.

Besides reasoning, choosing an accurate representation of time is challenging. In language, events are typically conceptualized as something that occurs, and they all have some duration. In the clinical domain, events can range from procedures to diseases to diagnoses, or to anything that the patient experiences. For simplicity, a point-based temporal logic is

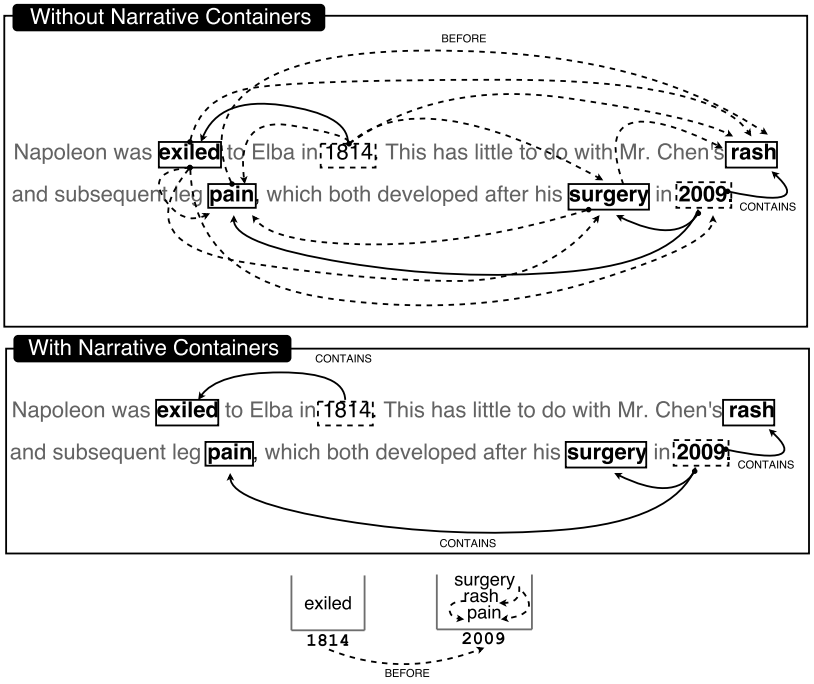

Fig. 2.1 Example temporal relation annotation with and without using narrative containers.

typically used to associate two periods of time. For example, given an event A ("surgery") and a time expression B ("tomorrow"), where A precedes B, we can infer a temporal relation BEFORE between A and B. Intuitively, we can also say that B comes AFTER A. The main problem with this temporal logic is that several temporal relations, which are not necessarily relevant to the reason about the situation described, can be identified within a text. Narrative containers are defined by Pustejovsky and Stubbs (2011) as an effort to reduce the scope of temporal relations between pairs of events and time expressions. As illustrated in Figure 2.1, narrative containers can be thought of as temporal buckets in which an event or a series of events may fall (Styler IV et al., 2014). They help visualize the temporal relations within a text and facilitate the identification of other temporal relation types. Until now, the only corpus annotated with this schema is limited to clinical texts.

2.1.1 Clinical TempEval challenges

Research on TIE has been instigated by Temporal Evaluation (TempEval) shared tasks that are focused on processing news article documents (UzZaman et al., 2013; Verhagen et al., 2007, 2010). However, in recent years, due to the high role of temporal reasoning in the interpretation of clinical narratives, the target domain has been shifted to the clinical domain. The resulting Clinical TempEval challenges (Bethard et al., 2017; Jones, 2015a,b) evaluate

systems on temporal information extraction from clinical notes and pathology reports from colon cancer patients, defining a series of sub-tasks that aim to identify temporal entities (EVENTS and TIMEX3: time expressions) and the temporal relations (TLINK) between them. Participating systems can choose to use raw text as input (phase 1) or they can use raw text with EVENT and TIMEX3 annotations (phase 2), in which case their task is to only to identify temporal relations. The temporal relation extraction track is further divided into two sub-tasks: (1) the identification of relations between events and the document creation time and (2) the identification of narrative container relations (TLINK:CONTAINS) between a directed pair (e_1, e_2). In this case, e_1 and e_2 are entities of either EVENT or TIMEX3 type. Clinical TempEval 2017 (Bethard et al., 2017) introduced a new aspect to the challenge, which still maintains the aforementioned sub-tasks but sets a new goal—to explore how well the systems trained in one medical domain perform on data from another. Such systems are trained on colon cancer data but are instead tested on brain cancer data.

Results of the systems participating in Clinical TempEval 2016 suggest that they perform well on time-entity identification tasks. Nevertheless, temporal relation extraction has proven to be the most difficult task. UTHealth (Lee et al., 2016), the best ranked system in Clinical TempEval 2016, showed a significant gap of 0.25 when compared to human performance[1] even when gold-standard entity annotations were provided. The improved task performance of recent works by Lin et al. (2016) and Leeuwenberg and Moens (2017) further enhanced the credibility of UTHealth's results, but the gap with respect to humans is still around 0.21. Regardless of the increase in the annotation agreement of temporal relations by relying on narrative containers, there is a consensus within the research community regarding the difficulties experienced in TIE. However, the reasons behind the skewed results between entity and temporal relation predictions still remain unclear.

2.1.2 Previous work

Until Clinical TempEval 2016, classic machine learning algorithms for classification such as conditional random fields, support vector machines (SVM) and logistic regression with a variety of features (e.g., lexical, syntactic and morphological) were the predominant approach to TIE (Jones, 2015a,b). In fact, the best performance was achieved by the UTHealth team (Lee et al., 2016) using an end-to-end system based on a linear and structural Hidden Markov Model (HMM)-SVM. Only a few teams tried a neural based method, including recurrent neural networks-based (RNN) models (Fries, 2016) and CNN-based models (Chikka, 2016; Li

[1]There are two scores for human performance: inter-annotator agreement and annotator-adjudicator agreement. We consider ann-adj as the upper bound performance since the models are trained on the adjudicated data, not on the individual annotator data (Bethard et al., 2017; Jones, 2015a,b)

and Huang, 2016). Furthermore, among those teams, only Chikka (2016) participated in the CONTAINS identification task, being around 0.30 below UTHealth's top performance.

Recent works by Lin et al. (2016), Dligach et al. (2017) and Leeuwenberg and Moens (2017) followed the settings of Clinical TempEval 2016 but they did not participate in the competition. Even though Leeuwenberg and Moens (2017) developed a new state-of-the-art model for temporal relation extraction, their results are still below human performance. Moreover, none of the aforementioned works provide a detailed discussion of *why* the current performance is so low and *how* the results on temporal relation extraction can be improved, save for Leeuwenberg and Moens, who in their first attempt on tackling this task on Clinical TempEval 2016 (Leeuwenberg and Moens, 2016), identified false negatives as their major problem.

Rather than a model's architecture or a dataset size, we believe that the complexity of temporal representation in natural language is likely to be the main cause of the low performance on temporal relation. *Tense* and *aspect* are the two grammatical means of expressing the notion of time in English, but little has been discussed about the latter in clinical texts. Furthermore, the focus of previous work on temporal relation extraction is set on narrative containers, relegating the identification of other temporal relation to a second place. However, we believe that the key is to look at the whole set of temporal types to achieve the ultimate goal of developing systems that can reason about time to automatically create a timeline of a patient's health care.

This study contributes to the current understanding of how temporal relations work in the clinical domain. It begins by illustrating how we adapt a general domain neural model for semantic relation extraction to temporal relation extraction from clinical text. It then analyzes the adopted system's overall performance, including the identification of narrative containers. Next, it discusses two major problems encountered when working with the narrative container's annotation schema, ending with a discussion on the necessary efforts needed to further improve the performance of the current state-of-the-art temporal relation extraction systems to perform on par with humans in terms of efficiency in completing the same tasks.

2.2 Methods

2.2.1 From relation extraction to temporal relation extraction

To determine the challenges in temporal relation extraction from clinical text, this study adapts a general domain relation extraction model. In NLP literature, the term "relation extraction" is a short form for "semantic relation extraction", which is the existing association

Table 2.1 Label distribution of pre-processed dataset for binary classification.

TLINK	Train	Test	Dev
CONTAINS	8,653	4,554	4,780
NONE	43,643	20,465	24,046
Total	52,296	25,019	28,826

Table 2.2 Label distribution of pre-processed dataset for multi-class classification.

TLINK	Train	Test	Dev
BEFORE	1,839	982	917
BEGINS-ON	717	363	298
CONTAINS	8,653	4,554	4,780
ENDS-ON	334	138	151
OVERLAP	2,388	1,186	1,582
NONE	43,643	20,465	24,046
Total	57,574	27,688	31,774

between the meaning of words, phrases, or sentences. Time concepts, such as duration and sequence, are embedded in a word's meaning (e.g., "bleeding" is an action that usually lasts for a moment and comes after another action like "cutting"). Therefore, semantic relations and temporal relations naturally overlap.

Relation extraction is a well-studied task in in NLP, where besides semantics, word sequence structures—such as recurrent neural networks (RNN) and linguistic features like the path of target words in the dependency tree—have shown to be effective (Xu et al., 2015). There is already a relation extraction model that integrates all of these elements—the end-to-end tree-based bidirectional long short-term memory-RNN model of Miwa and Bansal (2016). Due to its availability and state-of-the-art performance, we chose this model over Leeuwenberg and Moens's 2017 system. Moreover, we aim to take advantage of word sequence and dependency tree structures to further improve the performance on the Clinical TempEval relation extraction task. Given a sentence, Miwa and Bansal's 2016 three-layer model (i.e., embedding, sequence and dependency layers) jointly identifies entities and the relations between them. The model receives a sentence and an annotation file with a pair of terms as input and outputs the predicted relation type and directionality of the terms: (t_1, t_2) if t_1 is the source and t_2 the target, and (t_2, t_1) otherwise.

2.2.2 Experimental settings

Similar to Clinical TempEval 2016, we used the THYME corpus (Styler IV et al., 2014), a dataset of 600 clinical notes and pathology reports from colon cancer patients at the Mayo Clinic. The corpus is annotated at the document level and the identified entities are given a set of attributes depending on their type (i.e., *DocTimeRel*, *Type*, *Polarity*, *Degree*, *Contextual Modality* and *Contextual Aspect* for EVENTS and *Class* for TIMEX3). Temporal relation an-

Table 2.3 Performance of systems and humans on identifying CONTAINS relations. Our results come from five different random seeds.

System	P	R	F1
Lee et al., 2016 (UTHealth)	0.588	0.559	0.573
Lin et al., 2016	0.669	0.534	0.594
Our model	0.986	0.467	0.633
Human performance	-	-	0.817

notations specify source and target entities along with one of the following TLINK types: BEFORE, BEGINS-ON, CONTAINS, ENDS-ON and OVERLAP. Considering Miwa and Bansal's 2016 processes one sentence at a time, data was pre-processed to get sentence-level annotations. Any two EVENT/TIMEX3 can be a candidate pair. Therefore, all the entities in a sentence were used in generating all pair permutations as candidates. Pairs that did not have any temporal relations were then labeled as NONE (see Appendix A.1.1). The frequency of TLINKs in the THYME corpus is higher than the relations in the SemEval-2010 Task 8 dataset (Hendrickx et al., 2009), on which Miwa and Bansal's 2016 model was tested for relation classification. For this reason, we did not consider it necessary to extend the set of TLINKs to its transitive closure for data augmentation (i.e., A CONTAINS B $\wedge$ B CONTAINS C $\rightarrow$ A CONTAINS C). Table 2.1 and Table 2.2 detail the resulting datasets. In addition to the model's default Wikipedia word embeddings, we trained word vectors of 200 dimensions using word2vec (Mikolov et al., 2013) on a subset of PubMed2014.[2] PubMed2014 has 10,969,353 abstracts from 1,118,934 different journals. From those, we selected 634,813 abstracts from 38,677 journals related to Oncology and Gastroenterology. The MIMIC II clinical corpus (Saeed et al., 2011) is closer in genre to the THYME dataset, but due to its nature, one must get an application approval for its use. For simplicity purposes, we instead chose PubMed.

Next, we conducted four experiments at the intra-sentential level. The first experiment followed the Clinical TempEval 2016, focusing only on the identification of the CONTAINS type. The remaining experiments included all of the five annotated TLINKs. Further details of each of the experiments are given below:

I. **TLINK:CONTAINS binary classification:** In order to obtain results comparable to Lee et al. (2016), the best ranked system in Clinical TempEval 2016, we only considered TLINK:CONTAINS instances. The model chooses between CONTAINS and NONE relations.

[2]https://www.nlm.nih.gov/databases/download/pubmed_medline.html

II. **Multi-class classification with Wikipedia word embeddings:** To test the model in a real-world setting (i.e., a document that not only includes CONTAINS relations), we added the remaining pairs in the gold standard that have any of the other TLINK types to the train and test sets.

III. **Multi-class classification with PubMed word embeddings:** In addition to the previous setting (II), we used word embeddings trained on a subset of PubMed instead of the default word vectors trained on Wikipedia.

IV. **Multi-class classification with PubMed word embeddings and filtered negative examples:** Two extra difficulties of temporal relations are to ascertain whether an EVENT happened or not in a clinical context, and evaluate whether the said EVENT actually relates to the patient. For this reason, the THYME corpus differentiates "real" (*Contextual Modality:* ACTUAL or HEDGED) from "non-real" (*Contextual Modality:* HYPOTHETICAL or GENERIC) events. Real events cannot be related to non-real events. Therefore, in addition to the previous setting (III), we experimented removing a candidate pair whenever the e_1 contextual modality value[3] was ACTUAL or HEDGED and e_2 had HYPOTHETICAL or GENERIC modality, and vice versa.

2.3 Results

The evaluations were performed using the official Clinical TempEval scorer [4]. Table 2.3 shows performance on the CONTAINS identification task as a binary classification problem. The first row shows the top performance in Clinical TempEval 2016, while the second row is a result outside of the competition. We obtained an F1 score of 0.633, outperforming both UTHealth and Lin et al. (2016). Our model shows a high precision, but a lower recall than UTHealth; this can be attributed to the NONE relations prevalent in the dataset. Despite the recorded improvement, it is not possible to compare our system's performance with the current state-of-the-art set by Leeuwenberg and Moens's 2017, which was obtained using a multi-class classification approach. Table 2.4 reports our experimental results with the three multi-class classification settings presented in Section 2.2.2. Switching from binary classification to multi-class classification, we observe a significant drop in precision and a lower F1 score. This is expected because the classifier now has more TLINKS as options to choose from. Despite this change, our model outperforms both UTHealth and the state-of-the-art

[3]Note that entity attributes introduced at the beginning of this section were only used for pre-procesing, and not as features in our model.

[4]http://alt.qcri.org/semeval2016/task12/index.php?id=software

Table 2.4 Results of the three multi-class classification experiments and Leeuwenberg and Moens's 2017 Structured Perceptron (SP) best results on the THYME test set. The SP results were reproduced from the original paper. The results come from five different random seeds. FNE refers to filtered negative examples.

	Multi-class classification									
	Wikipedia word emb			*PubMed word emb*			*PubMed word emb + FNE*			*SP*
TLINK	P	R	F1	P	R	F1	P	R	F1	F1
BEFORE	0.696	0.183	0.289	0.704	0.196	0.306	0.683	0.213	**0.324**	0.294
BEGINS-ON	0.628	0.082	0.145	0.620	0.110	0.186	0.635	0.114	**0.194**	0.159
CONTAINS	0.907	0.468	0.617	0.904	0.471	**0.619**	0.900	0.472	0.618	0.608
ENDS-ON	0.525	0.093	0.157	0.656	0.122	0.204	0.637	0.115	0.194	**0.236**
OVERLAP	0.494	0.121	0.195	0.526	0.124	0.201	0.518	0.131	**0.209**	0.204
Macro-F1			0.281			0.303			**0.308**	0.300

model in terms of the F1 score of CONTAINS. Using PubMed word embeddings yielded the best F1 score for ENDS-ON and CONTAINS, and down-sampling negative examples on this setting improved the F1 score of BEFORE, ENDS-ON and OVERLAP. More details on the impact of using in-domain word embeddings and the FNE strategy are provided in Appendix A.2.

2.4 Discussions

Since this study did not change the architecture of Miwa and Bansal's 2016 model, the reader can, therefore, consult their study (Miwa and Bansal, 2016) for a detailed discussion on the system's performance on relation extraction. This section complements their discussion, which focuses on the linguistic characteristics of the dataset (clinical and temporal) that harms the system's performance.

2.4.1 Error analysis

Our error analysis focused on one-fourth of our experiments. Systems participating in the Clinical TempEval narrative container identification task only received credit for a pair of entities that they correctly identified the source, target, and the CONTAINS relation between them. Given this setting, we understand that even when using manual event and time annotations, the challenge is not only to predict the TLINK type but also the correct directionality of the entities. Therefore, part of our analysis aims to ascertain whether type classification or directionality identification is the most difficult task or if they are both equally problematic

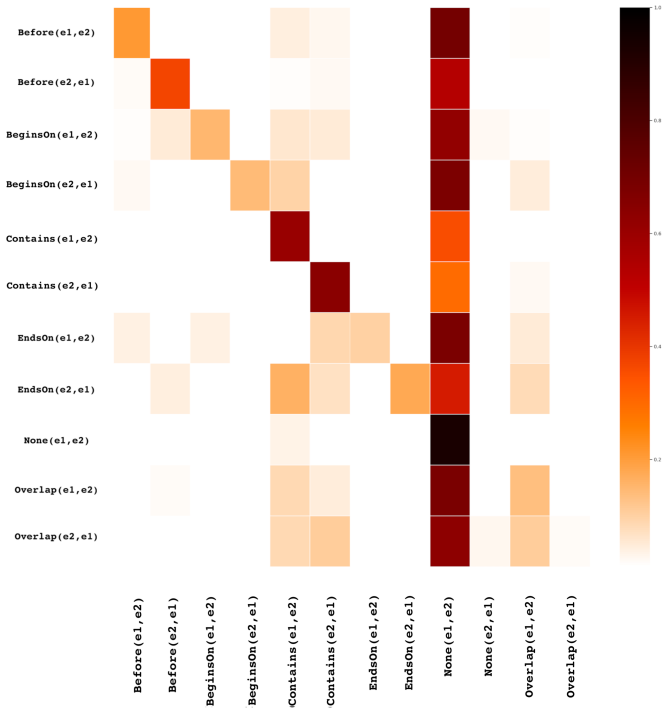

Fig. 2.2 Confusion matrix of our multi-class classification model with PubMed word embeddings on the dev set.

for the model. For this reason, we designed the confusion matrix using Miwa and Bansal's 2016 output instead of the Clinical TempEval 2016 script output. The confusion matrix on Figure 2.2 shows the results on the development set. Overall, due to the high number of negative instances, most of the false positives fall into the $None(e_1, e_2)$ category. This type of relation is the reason why the system shows high precision. Apart from this, we can identify the performance on OVERLAP as our system's main problem. The accuracy in both $Overlap(e_1, e_2)$ and $Overlap(e_2, e_1)$ is considerably low, with the latter being the lowest among all types (with 0.021). Not even the performance on $BeginsOn(e_2, e_1)$—with 0.14—is as low as $Overlap(e_2, e_1)$, although they have a similar number of instances (430 and 557, respectively). $Overlap(e_1, e_2)$, with 0.14, is comparable to $BeginsOn(e_2, e_1)$, despite having four times more instances (1,831 vs. 430). This explains why we focused our error analysis on OVERLAP.

From Figure 2, we can observe that $Overlap(e_1, e_2)$ is usually predicted as $Contains(e_1, e_2)$ and $Overlap(e_2, e_1)$ is predicted as $Contains(e_2, e_1)$. In both cases, the directionality of the entities was correct but the system failed to identify the appropriate temporal relation. For $Overlap(e_1, e_2)$, there were 112 sentences misclassified as $Contains(e_1, e_2)$, while in $Overlap(e_2, e_1)$ there were 32 $Contains(e_2, e_1)$ misclassifications. EVENT-EVENT pairs were the predominant type of pair in the former while TIMEX3-EVENT was for the latter, with 101 and 25 instances, respectively. We took all of the aforementioned misclassified sentences for supplementary examination and discuss the reason(s) for these errors in the following section.

2.4.2 Temporal relations and aspectual classes

Before proceeding further, it is important to understand the definition of OVERLAP and CONTAINS. Both temporal relations are closely related because they encompass the notion of two things happening at the same time. However, CONTAINS relations imply that the contained event (i.e., the target) occurs entirely within the temporal bounds of the event it is contained within (i.e., the source) while OVERLAP relations are those where containment is not entirely sure. Also, since e_1 OVERLAP e_2 means the same as e_2 OVERLAP e_1, OVERLAP is the only symmetrical TLINK type.

Time representation: Interval algebra and linguistics

Strictly speaking, every entity occupies time. An entity's time interval is crucial in understanding its temporal relation with respect to another entity, especially in the case of CONTAINS and OVERLAP relations where the end point of the target is key in determining whether there is complete containment or not. The temporal relations used by the THYME project rely on Allen's 1990 interval algebra, a precise way of expressing time periods using clear start and end points. By comparing those, we can easily indicate the position of two events on the timeline. However, the concept of time is widely discussed across disciplines and Allen's representation is just one among many others. In linguistics, the expression of time is understood because of two important grammatical systems: ***tense*** and ***aspect***.

Tense is used to locate the time of an event being talked about with respect to the time at which the speaker utters the sentence (i.e., speech time), while **aspect** is used to describe how a speaker views the contour of a situation (i.e., as beginning, continuation, or completion), independent of which position in time this situation occupies (Klein, 2013; Li et al., 2000). Therefore, when discussing a situation such as a patient having a surgery, we would use **past tense** if the surgery happened before speech time (*The patient had a surgery*), **future tense**

Category	C-Start	C-End	NC-Start	NC-End
Activity	+			+
Accomplishment	+	+		
Achievement	+*	+		
State			+	+

* *Start and end are so close to each other that this category considers no duration*

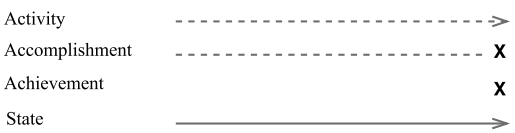

Fig. 2.3 Vendler's four-way classification. Arrows represent an indefinite time interval, solid lines indicate a homogeneous duration, and dashed lines indicate a dynamic duration. An X is used to represent a situation's natural end point. Abbreviations: C–Clear and NC–Not Clear.

if the surgery is about to happen (*The patient will have a surgery*), and **present tense** if the time of the surgery overlaps with the speech time (*The patient has a surgery*). Aspect, on the other hand, gives information about the surgery. The past tense in *The patient had a surgery* not only locates the surgery event before the speech time, but also conveys that the surgery was completed. This leads us to an important characteristic of tense and aspect—the boundaries between them are often not clear-cut (Li et al., 2000). The linguistic forms that express each of these notions tend to grammaticize into other categories (Bybee et al., 1994). In the case of English, the past tense form indicates past tense and perfective aspect (i.e., when a situation is reported in its entirety) simultaneously.

Having identified the functions of tense and aspect, we now focus on the latter. The study of aspect is commonly divided into *grammatical aspect* (also known as viewpoint aspect (Smith, 1983)) and the ***lexical aspect*** (also known as the situational aspect)[5]. Since temporal relations between events in the THYME project are thought in terms of their **start and endpoints**, the definition of the lexical aspect, which designates the internal temporal organization of the situation described by a verb, is particularly important to us (Klein, 2013). One of the best known and widely accepted aspect classification is that of Vendler, which distinguishes four categories of the inherent semantics of verb and verb phrases: *activities*, *accomplishments*, *achievements*, and *states* (Vendler, 1957). Figure 2.3 presents Vendler's classification using Andersen's 1990 schematization.

[5]For our discussion, we will use the term "aspect" to refer to lexical aspect.

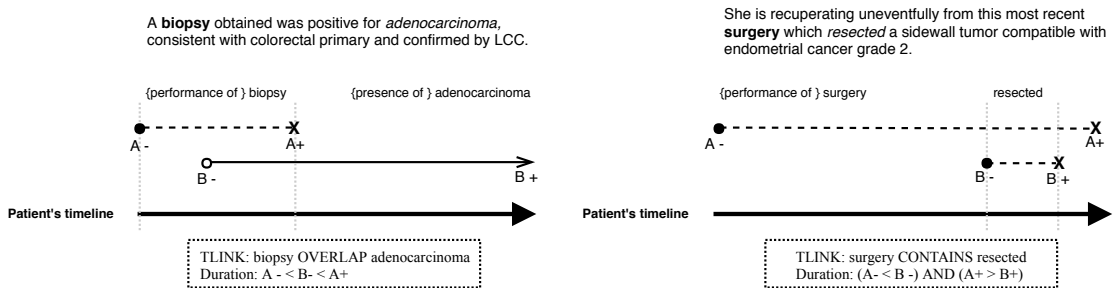

Fig. 2.4 Allen and Vendler's interval representation of OVERLAP and CONTAINS relations. A- / B- and A+ / B+ represent the start and end of an event, respectively. Filled-dots represent clear start points while an empty-dot represent a not-clear start point.

Aspectual classification of temporal entities

We expect that by categorizing the source and target entities of a relation as one of Vendler's types, the underlying reasoning for the TLINK classification will be simplified. For example, categories with **no clear end points** (such as *activities* and *states*) are more likely to overlap with those with **clear end points** (such as *accomplishments* and *achievements*). Figure 2.4 illustrates an OVERLAP and CONTAINS relations using Allen and Vendler's representation of time periods.

While analyzing OVERLAP relations that were mistaken for CONTAINS, we realized that only a few events were verbs. EVENTs in sentences 1, 3 and 9 in Table 2.5 are some examples of this ("invades," "seeking," and "moving"). This pointed out the necessity of discriminating between **verbal and non-verbal** events to understand how they are temporally related. Our observations suggest that when recognizing an entity semantic type (e.g., sign or symptoms, diseases, and procedures); it is imperative to consider the action associated to it. Therefore, procedures such as colonoscopy, biopsy, pathology, and surgery have to be *performed* (a dynamic verb with a natural end point–an *accomplishment*). Diseases such as adenocarcinoma and appendicitis are *present*, they exist, and consequently, they fall into the *state* category. This is also the case for signs or symptoms like nausea, fever, or discomfort. Following this line of reasoning, it is easier to differentiate an OVERLAP relation from CONTAINS in sentence 5 because we understand that nausea was present during the performance of the dialysis, but there is no enough information as regards to whether the nausea is still present or not. In other words, its end point is unclear. In the case of TIMEX3-EVENT pairs like those in sentences 8 to 10 in Table 2.5, the nature of the OVERLAP relation between the entities is due to the ambiguity of the time expressions combined with actions that we perceive as ongoing. For example, in sentence 9, the action of moving is an *activity* that is done indeterminably throughout the day as *multiple times a day* imply. On the other hand, in

Table 2.5 Sample of the analyzed misclassified sentences by our system. e_1 and e_2 are shown in bold and italics, respectively.

True relation	Predicted relation	Sentence
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	1. **Tumor** *invades* into the muscularis propria.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	2. Recurrent rectal **adenocarcinoma**, previously *resected* node-negative
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	3. Mr. Benefield is a pleasant 81-year-old male with resected colon **cancer** *seeking* treatment recommendations.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	4. Her **chemotherapy** was complicated by *angina* from the 5-FU which was treated with nitroglycerine, and her cardiac evaluation was negative.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	5. This morning, while at **dialysis**, she had *nausea*, fevers, and chills.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	6. Exploratory **surgery** with *appendicitis* many years ago.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	7. She was seen by a cardiologist in Idyllwild back in **April** when she was *hospitalized* and had an adenosine sestamibi scan after that hospitalization, but if surgery is contemplated I would wish her to be seen by cardiology.
$Overlap(e_2, e_1)$	$Contains(e_2, e_1)$	8. Does have some constipation with her iron supplementations but denies nausea, vomiting, abdominal distention, or worsening constipation, as she does have bowel **movements** *once every several days*.
$Overlap(e_2, e_1)$	$Contains(e_2, e_1)$	9. She is still **moving** her bowels *multiple times a day*.
$Overlap(e_2, e_1)$	$Contains(e_2, e_1)$	10. The patient **smokes** cigars *about once-a-month*.

sentence 7, there is a time expression with a definite time interval overlapping the patient's state of being hospitalized.

Styler IV et al. (2014) point out that several entities and other non-events are often interpreted in terms of their associated **eventive properties**. However, their discussion differs from ours in that they focus on how these properties define entities such as medications or disorders as an EVENT, rather than how the implicit interpretation of their eventuality (*taking a medication* or *having a disorder*) is necessary to relate two entities from a temporal perspective. They also introduce the **"contextual aspect,"** which is one of EVENT attributes, but their definition does not relate to the one used in linguistics. The contextual aspect attribute allows one of the three values,N/A, NOVEL, and INTERMITTENT, but as explained in the THYME guidelines, the N/A value simply represents an EVENT as neither NOVEL (i.e., new on the patient's timeline) nor INTERMITTENT (i.e., when there may be a series of smaller events within a single EVENT). The INTERMITTENT value can be useful in identifying an *activity* or an *accomplishment*, but as shown in Figure 2.5, just a small portion of EVENTs were

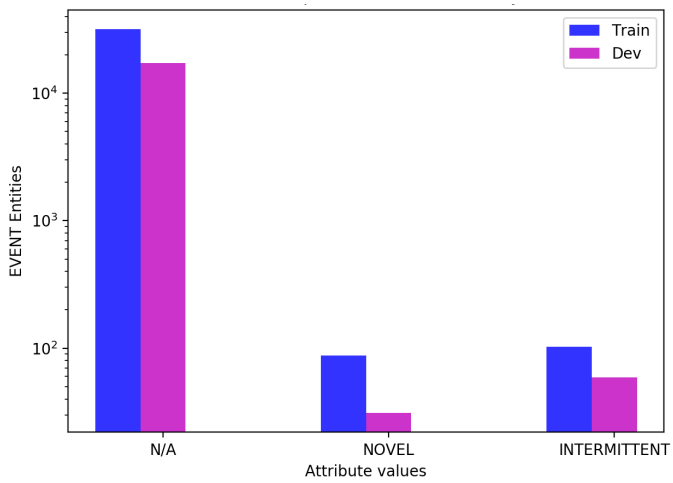

Fig. 2.5 CONTEXTUAL ASPECT attribute values by set

annotated with a value different from the default one. Moreover, **aspect is a property of verbs**, and our analysis insinuates that it is more common to find nouns as events.

The temporally locating of two events on a timeline requires a high level of reasoning that even humans can turn into a complicated task. All the aforementioned inferences for differentiating between two of the most frequent and most similar TLINK types (CONTAINS and OVERLAP) were done by heavily relying on the **internal constituency** of an event. Leveraging on aspectual type for temporal relation extraction is a promising approach that was explored by Costa and Branco (2012) on TempEval data, and our analysis implies that clinical data can also profit from it. However, this approach is limited since aspect is a property of verbs.

So far, we have been able to identify a **high similarity** of CONTAINS and OVERLAP relations as one of the reasons why these two types of TLINK are easily confused by our system, which did not pose much difficulties in identifying other TLINK types with a similar number of instances. This differs from what Styler IV et al. (2014) report for the annotator disagreement, which they say comes from different opinions about whether any two EVENTS require an explicit TLINK between them or an inferred one, *rather than what type of* TLINK *it would be (e.g.,* BEFORE *vs.* CONTAINS*)*. Our observations suggest that the main problem is not the amount of data available, but rather how temporal properties are encoded in language. The next section elaborates this point.

2.4.3 Temporality of nominal events

To deepen our understanding on the complexity of the temporal relation extraction task, we divided all OVERLAP and CONTAINS false negatives into the four possible pair types: EVENT-

Dev set: TLINK pairs				
TLINK	**E-E**	**T-T**	**E-T**	**T-E**
CONTAINS	149	6	2	37
OVERLAP	251	0	25	46
Total	400	6	27	83

Dev set: Event-Event pairs				
TLINK	**V-V**	**V-NV**	**NV-V**	**NV-NV**
CONTAINS	5	42	24	78
OVERLAP	3	54	24	170
Total	8	96	48	248

Table 2.6 Distribution of misclassified CONTAINS and OVERLAP pairs by type of TLINK (left) and EVENT type (right). Abbreviations: E–EVENT, T–TIMEX3, V–Verb and NV–Non-Verb

EVENT, TIMEX3-TIMEX3, EVENT-TIMEX3, and TIMEX3-EVENT. As shown in Table 2.6 (left), a significant amount of OVERLAP and CONTAINS links were EVENT-EVENT relations. Therefore, we looked further into this type of pairs, discriminating between verb (V) and non-verb (NV) events. Table 2.6 (right) shows the results in more detail.

As mentioned by Pustejovsky and Stubbs (2011) and further discussed in Styler IV et al. (2014), EVENT-EVENT pairings are a complex and vital component, particularly in clinical narratives, where doctors rely on shared domain knowledge and it is essential to read "between the lines." The distribution of verb/non-verb entities in Table 2.6 (right) indicates that most EVENT-EVENT misclassified pairings were either of NV-NV type or included a NV entity. This finding is of prime relevance to temporal reasoning since temporality is naturally encoded in verbs, expressing actions or events, while nouns are usually the person or thing doing or receiving that action (i.e., subject or object). Without a verb, the semantics of nouns hardly give a notion of time. Consider the following sentences:

(1) **Tumor** *invades* into the muscularis propria.

(2) Resected cecal **adenocarcinoma** with *resection* of liver metastasis.

In sentence 1, the NV entity "tumor" is annotated to overlap with the V entity "invades." Similarly, in sentence 2 "adenocarcinoma" is annotated to overlap with "resection." To understand how the time interval of these EVENT entities overlap, we inevitably look for an associated action to picture their duration, but the sole definition of *tumor* or *adenocarcinoma* does not provide us with that information. Clearly, it is not until we attribute these two nouns the property of *being present* (i.e., to exist) that we can think about a *state*, which has no inherent endpoint. This forced reasoning is not straightforward, even for humans. This depicts the fact that it could be even harder for computers to process.

Considering that verbs can be nominalized, we looked for nominalizations in our system's misclassified sentences. We observed some NV entities such as "consultation," "diagnosis," "discharge," "examination," and "resection," which derive from the verbs "to consult," "to diagnose," "to discharge," "to examine," and "to resect". However, it was more common to find NV entities like "cancer," "diabetes," "history," "anesthesia," and "dialysis," just to name a few. These entities are a good example of non-events being interpreted in terms of their (implicit) associated action. The THYME project defines an EVENT as *anything* relevant to the clinical timeline. This interpretation is broader than the one originally defined by Pustejovsky et al. (2005), where the term EVENT considers anything that *happens or occurs*, and is generally expressed by means of tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases. Consequently, the THYME project definition allows non-events, such as medications or disorders, to be annotated as EVENTS. While Styler IV et al. (2014) mentioned this, they did not show the frequency of nouns.

As we saw in previous examples, the time intervals of NV entities are more difficult to conceptualize, while V entities, such as "removed" or "improving," have their time properties morphologically encoded. Therefore, regardless of the low number of V-V relations, temporal information from verb predicates usually have more explicit hints; NV entities are more challenging and require more careful examination.

In Section 2.4.2, we pointed out the high similarity of OVERLAP and CONTAINS as one of the challenges of the temporal relation extraction task. Here we conclude that the high frequency of NV entities and the complexity of noun-noun relations is likely to be another reason why our system and previous works lag behind human performance. Not even the model of Miwa and Bansal (2016), which was designed to extract noun-noun relations, was able to handle the TLINKS in Clinical TempEval. As was noted earlier, the semantics of nouns are not enough to give the notion of an EVENT duration. This directly affects our system's performance.

We already introduced Vendler's aspectual classification and discussed how it helps to separate two extremely similar TLINKS. Unfortunately, this is not compatible with nominal predicates. In order to be able to use Vendler's topology in the clinical domain, the currently implicit associated actions to nominal EVENTS would have to be manually made explicit. Assuming that once identified, the verbs were classified as an *activity*, *accomplishment*, *achievement*, or *state*, this information can be used as a feature vector by our model. Alternatively, verb/non-verb entities distinction of EVENTS is the first step that can alleviate the incompatibility of aspect with nominals, and positively influence the temporal relation extraction task.

Table 2.7 Results of our multi-class classification experiments on the THYME test set. Our results come from five different random seeds. *Without None* refers to training without the None class.

	Multi-class classification											
	Wikipedia word emb			*Wikipedia without None*			*PubMed word emb*			*PubMed without None*		
TLINK	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BEFORE	0.696	0.183	0.289	0.573	0.461	**0.511**	0.704	0.196	0.306	0.567	0.466	**0.511**
BEGINS-ON	0.628	0.082	0.145	0.462	0.284	0.351	0.620	0.110	0.186	0.437	0.302	**0.357**
CONTAINS	0.907	0.468	0.617	0.810	0.647	**0.719**	0.904	0.471	0.619	0.814	0.641	0.717
ENDS-ON	0.525	0.093	0.157	0.551	0.325	**0.408**	0.656	0.122	0.204	0.462	0.306	0.368
OVERLAP	0.494	0.121	0.195	0.415	0.352	**0.381**	0.526	0.124	0.201	0.404	0.346	0.373
Macro-F1			0.281			**0.474**			0.303			0.465

2.4.4 Precision and recall imbalance

All the experiments introduced in Section 2.2.2 outperformed the best Clinical TempEval 2016 system and the state-of-the-art model as shown in Table 2.3 and Table 2.4. However, in all the settings, our model showed high performance, save for a recall lower than all of the systems presented in Table 3. Previously, we attributed this to the high number of negative instances in our dataset. As seen in Table 2.4, filtering some of the negative instances resulted in a slight increase in recall, but we were still unable to reach a better balance with precision. Moreover, our best recall score was still below the one achieved by Lee et al. (2016) and Lin et al. (2016). Consequently, we present an additional set of experiments to complement the multi-classification results in Table 2.4; i.e., training the model without the NONE class[6]. These results are shown in Table 2.7.

We can observe that the overall performance of our model remains the same. The classification of CONTAINS remains the highest, followed by BEFORE. The performance of the ENDS-ON and OVERLAP varies between the third and fourth best: ENDS-ON is the third best under the *Wikipedia without None* and *PubMed word embeddings*, while OVERLAP is under the *Wikipedia word embeddings* and *PubMed without None*. BEGINS-ON showed the lowest performance. Removing the NONE class resulted in a better balance between precision and recall, increasing the latter for all classes. Consequently, the F1 score increased as well. Despite this change, the OVERLAP still showed the second lowest performance.

[6]We did not remove the NONE class from the test set.

Class oversampling

At the beginning of these experiments, we assumed that the dataset size was not one of the underlying reasons of the low performance witnessed in previous works. Our error analysis indicated that low performance can be attributed to two reasons: the high similarity of OVERLAP with CONTAINS and the temporality of nominal events. Since we are using a neural model, it is natural to think that the more instances we have, the better the classification. We explored this assumption using four new experimental settings:

I. **Oversampling OVERLAP:** Given that OVERLAP has one of the lowest performance, even though it is the second most frequent class in the dataset, we balance OVERLAP with CONTAINS by random oversampling instances in the train set (ratio 1:1).

II. **Oversampling OVERLAP (without None):** Same setting as (I). The NONE class is removed for training but we maintain it for purposes o testing.

III. **Oversampling all the minority classes:** For better comparison, we balance all the minority classes (BEFORE, BEGINS-ON, ENDS-ON, and OVERLAP) with CONTAINS by random oversampling instances in the train set (ratio 1:1).

IV. **Oversampling all the minority classes (without None):** Same setting as (III). The NONE class is removed for training but we maintain it for purposes of testing.

The results are shown using default Wikipedia word embeddings and PubMed word embeddings in Table 2.8. Once again, we observe that the overall performance of our model remained the same. In the best case scenario, OVERLAP has the third best performance. However, it continues to show similar performance with the remaining minority classes. Similar to our first set of experiments, training with the NONE class results in high precision but low recall. In line with the results of Table 2.7, by removing the NONE class from the training set we get a better balance. Both oversampling techniques resulted in close Macro-F1 scores whenever we removed the NONE class from training and changed the default Wikipedia word embeddings for PubMed word embeddings (0.449 vs. 0.449 and 0.468 vs. 0.465, respectively). A McNemar's test McNemar (1947) on these results yields a p-value of 0.077 for *Wikipedia without None* and *PubMed without None* oversampling OVERLAP and a p-value of 0.466 when oversampling all minority classes. This is not significant at the 0.05 alpha level. The reason might be the same as why in-domain embeddings had a limited improvement in Section 2.2.2 multi-class classification experiments (see Appendix A.2.1).

Table 2.8 Results of our multi-class classification experiments on the THYME test set. The first results are results obtained when oversampling only OVERLAP, and the subsequent results are obtained when oversampling BEFORE, BEGINS-ON, ENDS-ON, and OVERLAP.

	Multi-class classification oversampling OVERLAP											
	Wikipedia word emb			*Wikipedia without None*			*PubMed word emb*			*PubMed without None*		
TLINK	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BEFORE	0.686	0.165	0.266	0.592	0.423	0.494	0.679	0.155	0.252	0.571	0.447	**0.501**
BEGINS-ON	0.593	0.090	0.157	0.478	0.253	**0.331**	0.727	0.082	0.148	0.416	0.242	0.306
CONTAINS	0.900	0.461	0.610	0.810	0.631	**0.709**	0.909	0.448	0.601	0.814	0.618	0.703
ENDS-ON	0.583	0.046	0.086	0.500	0.265	0.346	0.684	0.086	0.153	0.490	0.318	**0.386**
OVERLAP	0.365	0.164	0.227	0.355	0.373	**0.363**	0.373	0.188	0.250	0.344	0.360	0.351
Macro-F1			0.269			**0.449**			0.281			**0.449**

	Multi-class classification oversampling all minority classes											
	Wikipedia word emb			*Wikipedia without None*			*PubMed word emb*			*PubMed without None*		
TLINK	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BEFORE	0.629	0.287	0.394	0.549	0.446	0.492	0.622	0.287	0.393	0.532	0.469	**0.499**
BEGINS-ON	0.466	0.209	0.288	0.375	0.340	**0.357**	0.392	0.201	0.266	0.325	0.312	0.318
CONTAINS	0.912	0.449	0.602	0.827	0.614	**0.705**	0.923	0.441	0.597	0.836	0.603	0.701
ENDS-ON	0.448	0.371	0.406	0.424	0.424	0.424	0.456	0.344	0.392	0.415	0.450	**0.432**
OVERLAP	0.409	0.165	0.235	0.368	0.354	0.361	0.410	0.164	0.234	0.379	0.367	**0.373**
Macro-F1			0.385			**0.468**			0.376			0.465

2.5 Conclusions

Clinical language processing represents a special challenge to NLP systems. The structure of clinical texts range from telegraphic constructions to long utterances describing a patient's condition or a suggested diagnosis. The high use of domain knowledge to infer temporal relations between events does not make this task any easier. A doctor naturally interprets adenocarcinoma (a type of cancer) as an abnormal, uncontrolled and *progressive* growth of tissue, which temporally speaking is and should be thought as an ongoing process unless explicitly qualified (*"We resected the adenocarcinoma, and since margins were clear, we can say it is gone"*). This is a non-trivial task for a computer even when relying on context information.

There have been several attempts on tackling temporal relation extraction from clinical text, mostly led by the Clinical TempEval challenges. However, the results are still far from human performance and there is little information about the underlying reasons. This encouraged our work to adapt a state-of-the-art system and do a detailed error analysis, which pointed out that one of the major challenges is how to handle the eventive properties

of nominals—the predominant type of events on the most frequent type of pairs (EVENT-EVENT).

Existing knowledge bases, such as the Unified Medical Language System's (UMLS) Metathesaurus help to classify entities into semantic types like *Therapeutic or Preventive procedure*, *Sign or Symptom* or *Disease or Syndrome*. However, the associated events and actions cannot be found in this or any other knowledge base. Therefore, we hypothesize that a resource containing aspectual information of the actions associated to common nominals, such as procedures or diseases, can further improve temporal relation extraction in the clinical domain. Since this will require manual annotation effort from annotators with linguistic and clinical knowledge, we first plan to analyze further EVENT-EVENT relations by differentiating events as verbal and non-verbal events.

Chapter 3

Evaluation of image descriptions for commonsense reasoning in machine reading comprehension

In this chapter, we explore the performance on a commonsense-based machine reading comprehension task. As a first step, we present a set of preliminary experiments to evaluate the existence of commonsense knowledge in image descriptions.

3.1 Introduction

The recent advances achieved by large neural language models (LMs), such as BERT (Devlin et al., 2018), in natural language understanding tasks like question answering (Rajpurkar et al., 2016) and machine reading comprehension (Lai et al., 2017) are, beyond any doubt, one of the most important accomplishments of modern natural language processing (NLP). These advances suggest that a LM can match a human's stack of knowledge by training on a large text corpora like Wikipedia. Consequently, it has been assumed that through this method, LMs can also acquire some degree of *commonsense* knowledge. It is difficult to find a unique definition, but we can think of *common sense* as something we expect other people to know and regard as obvious (Minsky, 2007). However, when communicating, people tend not to provide information which is obvious or extraneous (as cited in Gordon and Van Durme (2013)). If common sense is something obvious, and therefore less likely to be reported, what LMs can learn from text is already being limited. Liu and Singh (2004) and more recently Rashkin et al. (2018) and Sap et al. (2019) have tried to alleviate this problem by collecting crowdsourced annotations of commonsense knowledge around frequent phrasal events (e.g.,

PERSONX EATS PASTA FOR DINNER, PERSONX MAKES PERSONY'S COFFEE) extracted from stories and books. From our perspective, the main limitation of this approach is that even if we ask annotators to make explicit information that they will usually omit for being too obvious, the set of commonsense facts about the human world is too large to be listed. Then, what other options are there?

As the name suggests, common sense[1] is related to *perception*, which the Oxford English Dictionary defines as the ability of becoming aware of something through our senses: SIGHT *(e.g., the sky is blue)*, HEARING *(e.g., a dog barks)*, SMELL *(e.g., trash stinks)*, TASTE *(e.g., strawberries are sweet)*, and TOUCH *(e.g., fire is hot)*. Among those, vision (i.e., sight) is one of the primary modalities for humans to learn and reason about the world (Sadeghi et al., 2015). Therefore, we hypothesize that annotations of visual input, like images, are an option to learn about the world without actually experiencing it. This chapter explores to what extent the textual descriptions of images about real-world scenes are sufficient to learn common sense about different human daily situations. To this end, we use a large-scale image dataset as knowledge base to improve the performance of a pre-trained LM on a commonsense machine reading comprehension task.

We find that by using image descriptions, the model is able to answer some questions about common properties and locations of objects that it previously answered incorrectly. If we prove our hypothesis to be true we would have an alternative to the expensive (in terms of time) and limited (in terms of coverage) crowdsourced-commonsense acquisition approach.

3.2 Related work

Knowledge extraction. Previous works have already recognized the rich content of computer vision datasets and investigated its benefits for commonsense knowledge extraction. For instance, Yatskar et al. (2016) and Mukuze et al. (2018) derived 16K commonsense relations and 2,000 verb/location pairs (e.g., *holds*(dining-table, cutlery), *eat*/*restaurant*) from the annotations included in the Microsoft Common Objects in Context dataset (Lin et al., 2014) (MS-COCO). However, they only focused on physical commonsense. A more recent trend is to query LMs for commonsense facts. While a robust LM like BERT has shown a strong performance retrieving commonsense knowledge at a similar level to factual knowledge (Petroni et al., 2019), this seems to happen only when that knowledge is explicitly written down (Forbes et al., 2019).

Machine reading comprehension (MRC). MRC has long been the preferred task to evaluate a machine's understanding of language through questions about a given text. The current

[1] Latin *sensus* (perception, capability of feeling, ability to percieve)

most challenging datasets such as Visual Question Answering (Goyal et al., 2017), NarrativeQA (Kočiskỳ et al., 2018), MCScript (Ostermann et al., 2018, 2019), CommonsenseQA (Talmor et al., 2018), Visual Commonsense Reasoning (Zellers et al., 2019) and CosmosQA (Huang et al., 2019) were designed to be solvable only by using both context (written or visual) and background knowledge. In all of these datasets, no system has been able to reach the upper bound set by humans. This emphasizes the need to find appropriate sources for systems to equal human knowledge.

This work lies in the intersection of these two directions. We aim to use computer vision datasets for broad commonsense knowledge acquisition. As a first step, we explore whether visual text from images provides the implicit knowledge needed to answer questions about an MRC text. Ours is an ongoing attempt to emulate the success of multi-modal information in VQA and VCR on a MRC task.

3.3 Approach

We evaluate image descriptions through a MRC task for which commonsense knowledge is required, and assume that answering a question incorrectly means the reader lacks such knowledge. Most of what humans consider obvious about the world is learned from experience, and we believe there is a fair amount of them written down in an image's description. We will test this idea by using image descriptions as external knowledge. Out of the different types of common sense, the text passages in the selected MRC dataset focus on *script knowledge* (Schank and Abelson, 2013), which covers everyday *scenarios* like BRUSHING TEETH, as well as the *participants* (persons and objects) and the *events* that take place during them. Since scenarios represent activities that we do on a regular basis, we expect to find images of it. Ideally, for each passage, we would automatically query an image dataset to retrieve descriptions related to what the passage is about. Retrieval is a key step in our approach and for the time being, such process was done manually so we can focus on the image's description content rather than in the retrieval process itself.

There is a considerable number of crowdsourced image datasets whose image descriptions are available, which means they can be collected (and extended, if needed) for a reasonable cost. The motivation behind our approach is that once such descriptions are proven to contain useful commonsense knowledge that it is not easily obtained from text data, one can think of extending the description collection.

3.4 Experiments

3.4.1 Data

Image dataset. Visual Genome (Krishna et al., 2017) is a large-scale collection of non-iconic, real-world images with dense captions for multiple objects and regions in a single image. Each of the 108K images in the dataset has an average of 50 region descriptions of 1 to 16 words. To use this dataset as a knowledge base, we first used BERT-sentence embeddings (Reimers and Gurevych, 2019) to embedded all of the region descriptions and then created a semantic search index using FAISS (Johnson et al., 2017). When querying the index, we retrieved the top 50 results.

Reading comprehension dataset. MCScript2.0 is a dataset with stories about 200 everyday scenarios. Each instance has a text passage paired with a set of questions, which in turn have two answer candidates (one correct and one incorrect). In total, MCScript2.0 has 19,821 questions, out of which 9,935 are commonsense questions that require script knowledge. Figure 3.1 shows a diagram with the distribution of questions. We split the dataset into train, dev and test sets as in (Ostermann et al., 2019). The train set is used as it is. However, for evaluation, we worked with a subset of 56 and 81 questions from the original dev and test sets, respectively (more details of this in the next section). The subsets include instances with passages about 15 out of the 200 scenarios. For each instance, we took all of its commonsense questions and further selected those in which the necessary commonsense knowledge might be present in one (or more) image descriptions. An example is shown in Figure 3.2.

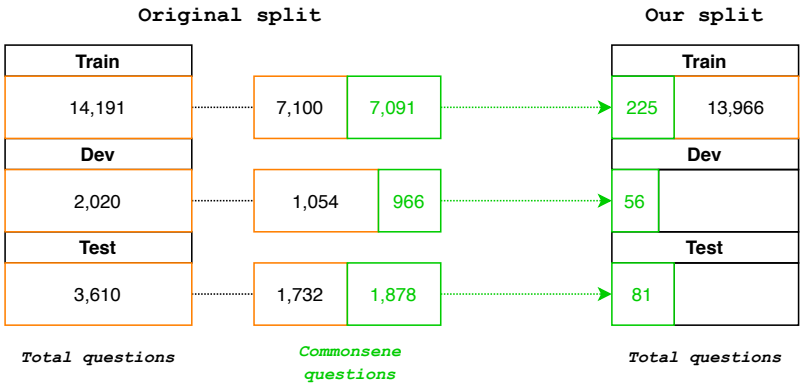

Fig. 3.1 Visualization of MCScript2.0 original data split and our data split.

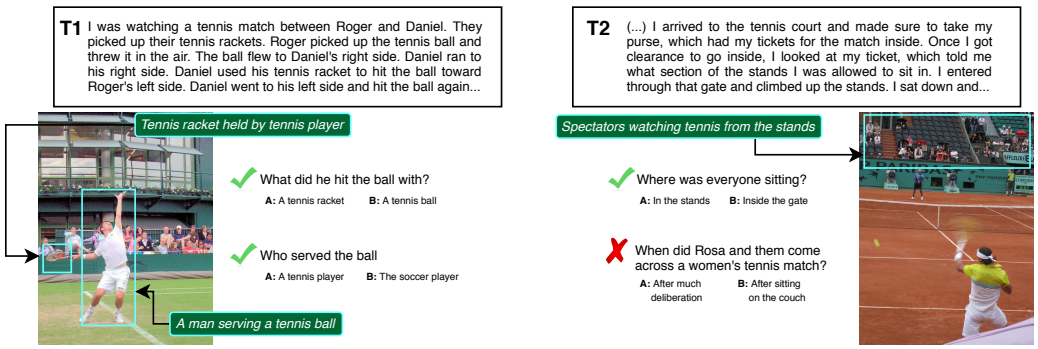

Fig. 3.2 Example of three selected and one removed commonsense questions from two MCScript2.0 instances.

3.4.2 Models

BERT (Baseline). We fine-tuned a vanilla-BERT with the following input configuration: the question and one of its answer candidates are appended to segment one and the text passage is appended to segment two. Therefore, we have two inputs per instance. To help BERT differentiate between the question and answer-candidate tokens, we used a special separator token[2]. The maximum sequence length was set to 384. Figure 3.3 shows how we build the input representation of two MCScript2.0 questions. The question and answer candidate A in segment one, and the text passage in segment two. Similarly, there is a second input representation with the question and answer candidate B in segment one, and the text passage in segment two. BERT computes a softmax over the two choices to predict the correct answer candidate. Visually Enhanced BERT (detailed below) builds the input in a similar way. The difference is that the manually selected region descriptions are appended at the beginning of the text passage. The number of tokens in the text passage increases, but the input configuration remains the same.

We trained the model up to 5 epochs with a learning rate (Adam) of 5e-5 and a training batch size of 8 using 3 different random seeds.

Visually Enhanced BERT. As introduced in Section 3.3, we hypothesize there is commonsense knowledge present in image descriptions. This model aims to improve on the baseline by using region descriptions from Visual Genome to answer those questions were BERT was wrong. We will refer to these questions as the *unanswerable questions* set. All of them were manually inspected to identify the scenario they are about. As shown in Figure 3.4, the scenario name is used to query our Visual Genome index. If the results do not contain information about the scenario's events or participants, we refined the query using keywords

[2]We used '[unused00]' as the special separator token, which is included in BERT's vocabulary

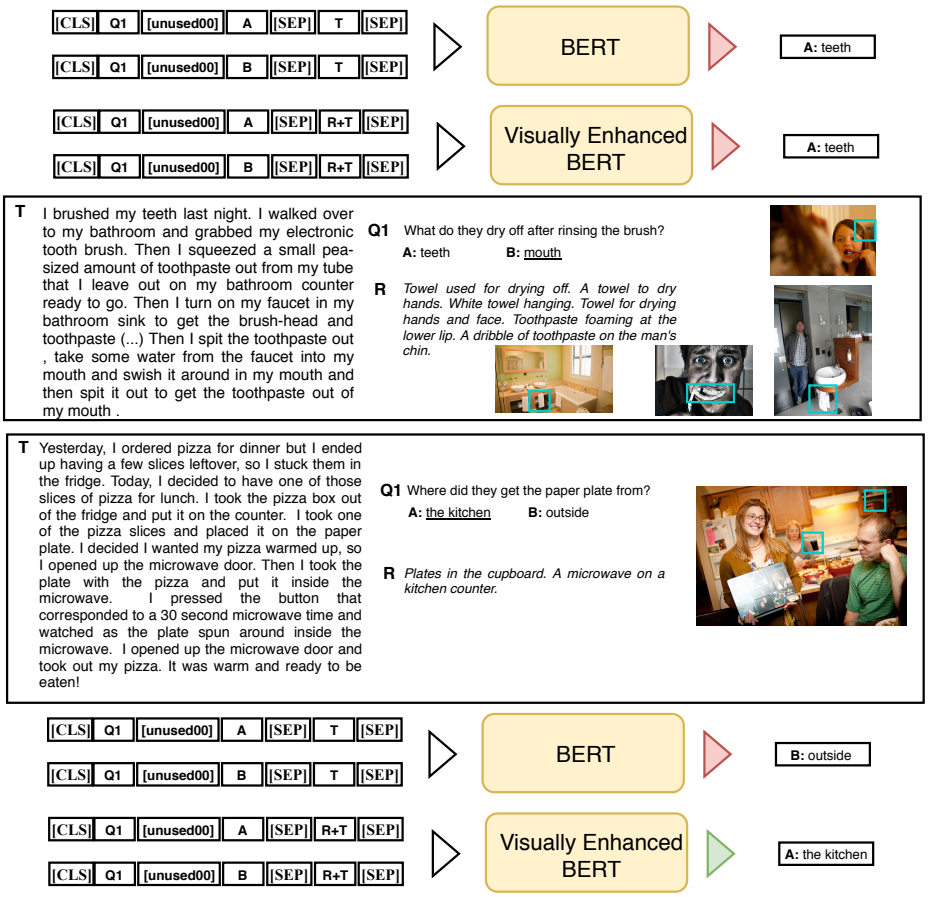

Fig. 3.3 Two input/output examples. In the top example, region descriptions were not helpful to chose the correct answer candidate. In the bottom example, they were.

from the question (e.g., querying “going fishing” returns no results mentioning “rod”, a new query would be “going fishing rod”). To be careful not to exceed BERT’s sequence length, we selected a maximum of 6 region descriptions from the results and concatenated them at the beginning of the given question’s text passage. Finally, we fine-tuned the model just as we did with the baseline model.

The whole retrieval process was done manually, which did not represent much of a problem for the dev and test subsets. However, it would be time-consuming to follow this approach with the train set. We fine-tuned on the complete train data, but we limited the use of image descriptions to 225 train questions that were selected in the same way as the dev and test subsets.

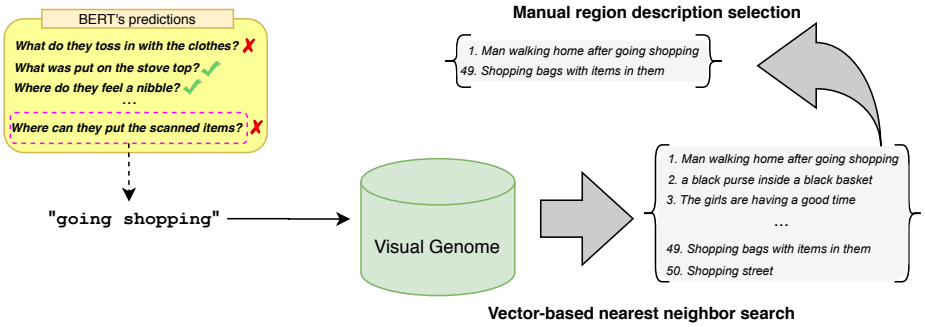

Fig. 3.4 Retrieval process for one of the questions BERT answered incorrectly. Identifying the GOING SHOPPING scenario, querying Visual Genome and selecting the most related region descriptions to the scenario was manually done.

3.5 Results

For most of the questions in the unanswerable set, we did find related region descriptions. Figure 3.5 shows some of the images retrieved and the regions that matched what the question is asking. Besides its size, one of the main advantages of Visual Genome annotations is that they cover several regions that compose the scene in an image. Thanks to this, we were able to find region descriptions that not only mention an object (e.g., *a towel, scissors, a dollar-bill*), but also add a description of how the object can be used (e.g., *towel used for drying off, scissors for cutting string*) or what does it represent (e.g., *five dollar tip on table*). This suggests that our hypothesis mentioned in Section 3.1 about annotations of visual input might be correct. As shown in Table 3.1, region descriptions helped BERT

	Dev	**Test**
Model	Commonsense	Commonsense
fine-tuned BERT (base-uncased)	.780	.732
Visually Enhanced BERT	.857	.749

Table 3.1 Accuracy of BERT baseline and our manually visually enhanced BERT in both MCScript2.0 development and test sets. The results come from three different random seeds.

to achieve a better accuracy. If our hypothesis is true, the improvement should come from correctly answering questions from the unanswerable set. This was true for those related to affordances.[3] Some examples of questions that became answerable for Visually Enhanced BERT are *What did they toss in with the clothes?*, and *What do they cut out the pieces with?*. Another type of question BERT initially had problems answering required commonsense

[3] An object's properties that show the possible actions users can make with it.

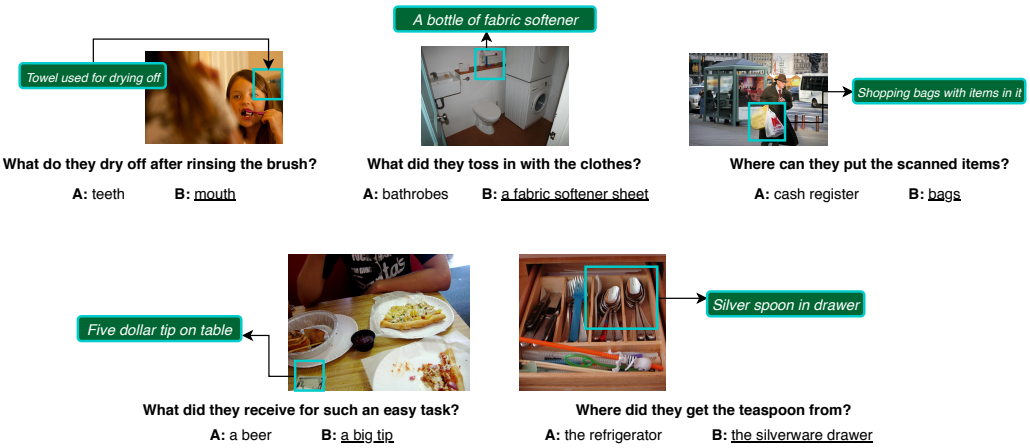

Fig. 3.5 Examples of questions from the unanswerable set and one of the manually selected region descriptions from Visual Genome.

knowledge about an object's location. Some examples of those questions are *Where did they get the teaspoon from?* (Answer: the silverware drawer) and *Where did they get the paper plate from?* (Answer: the kitchen). Our results suggest that region descriptions were more beneficial to these type of questions, since they were no longer unanswerable for Visually Enhanced BERT. However, there were cases in which we could not see an improvement. Questions like *What did they receive for such an easy task?* (Answer: big tip) and *What does a list keep them on?* (Answer: budget) do require commonsense knowledge about the SERVING A DRINK and GOING SHOPPING scenarios, but the concept that needs to be understood is too abstract. Even though we found region descriptions that match the correct answer candidate (e.g., *Five dollar tip on table. Tip on the table.*), these type of questions remained unanswerable for Visually Enhanced BERT.

In a classic reading comprehension task, word matching usually helps to find the correct answer. However, MCScript2.0 evaluates beyond mere understanding of the text and as such, it was designed to be robust against it. Out of the 56 questions in our dev set, we observed that the number of times a passage mentions the correct and the incorrect answer candidates is similar (42 and 36, respectively) and in either case this seemed to have influenced BERT's predictions. This stayed roughly the same after we appended the region descriptions.

3.6 Conclusion

Pre-trained large LMs have significantly closed the gap between human and computer performance in a wide range of tasks, but the commonsense knowledge they capture is still limited. In this chapter, we presented a controlled experimental setup to explore the plausibility of

acquiring commonsense knowledge from dense image descriptions. Our preliminary results on a commonsense-MRC task suggest that such descriptions contain simple but valuable information that humans naturally build through experiencing the world. In future work, our aim is to automate the retrieval process and to extend the evaluation subset.

Chapter 4

Comparing the content of two text sources and their impact on commonsense machine reading comprehension

In Chapter 3, we hypothesize that the textual descriptions generated by referring to a visual input (an image) contain a fair amount of commonsense knowledge about every-day scenarios. For example, for a common activity like BRUSHING TEETH, we could find an image description mentioning the typical location of a toothbrush (*a toothbrush in a cup holder*), how do we use toothpaste (*a man putting toothpaste on a toothbrush*) or where do we usually brush our teeth (*scene taking place in bathroom*). We automatically queried an index with region descriptions, but we manually selected the ones that were more related to a given scenario. Since manual selection is a time-consuming process, we experimented with a small subset of MCScript2.0. In this chapter, we fully automatize the retrieval process and extend the evaluation set. In addition to the image description's index, we also query a well-known commonsense knowledge base and compare the results using each one of the sources as external knowledge on MCScript2.0's machine reading comprehension task.

4.1 Introduction

Reading comprehension is considered a complex and highly demanding cognitive task that involves the simultaneous process of extracting and constructing meaning (Kintsch and Walter Kintsch, 1998). This task is commonly used to evaluate a person's language competence

T	(...) We put our ingredients together to make sure they were at the right temperature, preheated the oven, and pulled out the proper utensils. We then prepared the batter using eggs and some other materials we purchased and then poured them into a pan. After baking the cake in the oven for the time the recipe told us to, we then double checked to make sure it was done by pushing a knife into the center. We saw some crumbs sticking to the knife when we pulled it out so we knew it was ready to eat !
Q1	*When did they put the pan in the oven and bake it according to the instructions?*
	After eating the cake. ✗
	After mixing the batter. ✓
Q2	*What did they put in the oven?*
	The cake mix. ✓
	Utensils. ✗

Fig. 4.1 Example text fragment from MCScript2.0

and likewise, it is used in NLP to test a system's language understanding. The core elements of reading comprehension are a *text passage* and a set of *questions*. With the aim of making the task a little bit less complicated than it already is, machine reading comprehension (MRC) usually comes with a third element: a set of *answer candidate choices*. Thus, whenever we talk about MRC we usually refer to multiple-choice MRC. Given these 3 elements, a good reader should be able to find the association between them. Consider the sample text passage from MCScript2.0 shown in Figure 4.1.

The two questions can be directly answered (i.e., without referring to the passage) by someone who is familiar with a the COOKING scenario. For someone who is not, keywords in the text like *temperature*, *preheated* and a phrase like *pushing a knife into the center* give some hints about the required knowledge to choose *the cake mix* as the correct answer candidate over *utensils*.

Ostermann et al. (2019)'s human evaluation of MCScript2.0 showed that humans answer the questions on this dataset with an average accuracy of 97.4%. What is it that humans do to comprehend a text? According to Zwaan (1999), comprehension is first and foremost the construction of a mental representation of what the text is about. His *situation model* theory states that we do not rely so much on the structure of the text, we construct mental representations of the people, objects, locations, events, and actions described in it. In other

words, we could say that we create a *mental image* about the text. In the case of the passage shown in Figure 4.1, to answer the question *What did they put in the oven* we imagine an oven and then, we imagine the cake mix and the kitchen utensils in it. We immediately identify that the latter situation would just not happen, leaving *the cake mix* as the correct choice.

Under the aforementioned background, our intuition is that a large-scale image dataset is likely to have images with different scenes of everyday scenarios and their textual descriptions will contain knowledge about the characteristics of related object and actions. However, this typical, commonsense knowledge can also be found in a knowledge base. Recognizing that there is information that is not available to computers since they cannot experience the world, the NLP community has made an effort to annotate a series of commonsense facts. Both image descriptions and the entries of a knowledge base are text sources, but they differ in how they were created: the former was generated by an annotator that was looking at an image and it was asked to describe it, while in the latter an annotator is prompted with a text event like *cooking* and it is asked to annotate facts related to it. Do these sources also differ in the knowledge they provide? The goal of this chapter is to find out if there is knowledge contained in one source that cannot be found in the other.

The experiments in this chapter improve the approach presented in Section 3.3 as follows:

I. **Automatic retrieval process**: In Chapter 3 we evaluated the content of image descriptions using commonsense MRC as a downstream task. We hypothesized that there is commonsense knowledge in image descriptions and that if we used them as external knowledge, our model will improve its performance on answering commonsense questions. Those image descriptions were manually selected. Here, we present an alternative to automatically retrieve image descriptions.

II. **Bigger evaluation set**: Due to the manual retrieval of image descriptions, our experiments in Chapter 3 were limited to a small evaluation subset of MCScript2.0 data. We extend the evaluation set to draw more meaningful conclusions.

4.2 Automating external knowledge retrieval

Each text on MCScript2.0 describes a situation about an everyday scenario like BRUSHING TEETH. Thus, we need to retrieve information related to the scenario the text passage is talking about. For a given passage, question and its answer candidates, we can break down the comprehension process in two:

I. Associating an answer candidate with the question

II. Associating an answer candidate with the text passage

In the following subsections, we will detail two simple ways of querying an index taking into consideration the aforementioned comprehension processes.

4.2.1 Question-answer retrieval

As implied in the name, this setting only considers the *question* and an *answer candidate*. Using these two elements, we retrieve information from a semantic search index as follows:

I. We **query** the index using the embedding representation of the question.

II. We **re-rank** the top 50 results based on their cosine similarity with the given answer candidate.

III. After re-ranking the contents in the result set, we **keep** the top 10 results.

4.2.2 Question-passage-answer retrieval

In this setting we consider the three core elements of MRC: *passage*, *question* and *answer candidate*. The process is essentially the same as the one described in section 4.2.1:

I. **Query** the index using the embedding representation of the question.

   (a) **Re-rank** the top 50 results based on their cosine similarity with the given answer candidate.

   (b) **Cluster** the top 20 results in 5 clusters.

   (c) **Keep** one result for each of the clusters.

II. **Query** the index using the embedding representation of the text passage.[1]

   (a) **Re-rank** the top 50 results based on their cosine similarity with the given answer candidate.

   (b) **Cluster** the top 20 results in 5 clusters.

   (c) **Keep** one result for each of the clusters.

The final top 10 results is the concatenation of I.(c) and II.(c).

[1] Sum of each sentence's sentence embedding.

4.3 Experiments

4.3.1 Indexes

Image descriptions: Descriptions from Visual Genome (Krishna et al., 2017), a large-scale collection of non-iconic, real-world images with dense captions for multiple objects and regions in a single image. Each of the 108K images in the dataset has an average of 50 region descriptions of 1 to 16 words.

General commonsense knowledge: English surface text of ConceptNet (Liu and Singh, 2004) triples.[2] A triple in ConceptNet has a *start edge*, *end edge* and a pre-defined *relation* between them (e.g., *IsA*, *LocatedNear*, etc). For example, the surface text of triple *LocatedNear(pillow case, bed)* is *pillow case is typically near bed*.

Both indexes are created by first embedding each of their entries (region descriptions in Visual Genome and triples in ConceptNet) as sentence embeddings using SBERT (Reimers and Gurevych, 2019). The embeddings are indexed using FAISS (Johnson et al., 2017). The Visual Genome index has 5,408,689 entries and the ConceptNet index has 3,895,394 entries.

4.3.2 Data

In Section 3.4.1 we described a process to select a subset of MCScript2.0 data. Since the image descriptions were going to be manually selected, we had to limit the evaluation set to a small subset of questions. In this chapter, we automated the retrieval process, which allows us to test on a bigger subset. Figure 4.2 shows how did we split the data and Table 4.1 detail the resulting datasets.

	Train	**Dev**	**Test**
Total instances	2,250	250	60
Commonsense questions	6,351	740	137
Total questions	12,732	1,459	137

Table 4.1 Distribution of MCScript2.0 instances and questions on each data split.

This time, we completely focus on MCScript2.0 original TRAIN split. We left the original DEV and TEST splits unseen for future work on the MCScript2.0 dataset. The instances in the original TRAIN split were divided in 10 folds. We use 9 of those folds as our new training

[2]https://github.com/commonsense/conceptnet5/wiki/Downloads

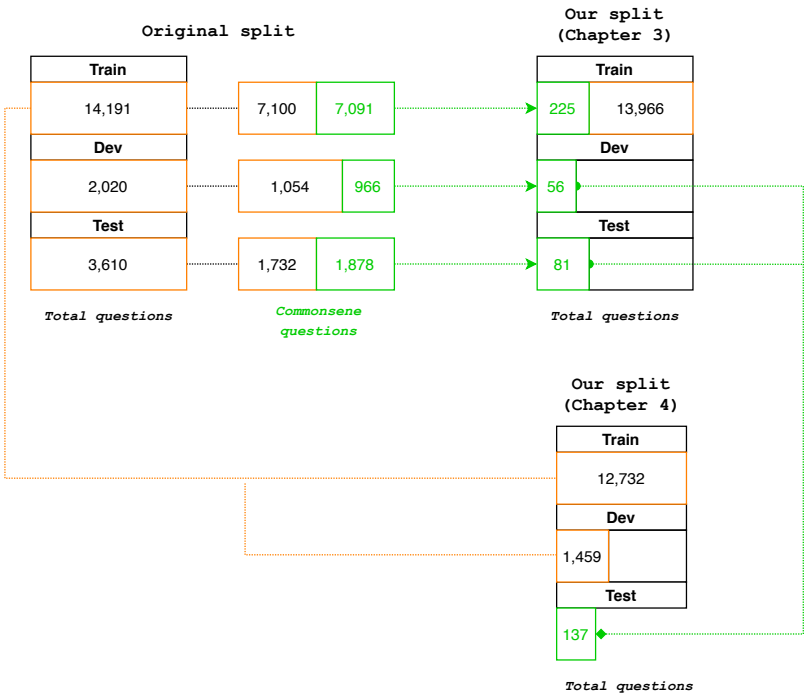

Fig. 4.2 Visualization of MCScript2.0 original data split. In the right, a visualization of how we derived the data used in Chapter 3. Below, the data split used in the current chapter.

set, the remaining fold as evaluation (dev) set and we merged the DEV and TEST small subsets from Chapter 3 to be our new test set.

4.3.3 Models

We experiment using BERT in the same way we did in Chapter 3, using BERT without any type of external information as baseline and adding the queried information from Visual Genome and ConceptNet to BERT's input for comparison. Below is an overview of our three models:

I. **BERT (baseline)**: BERT fine-tuned on our MCScript2.0 subset. Given that each question in MCScript2.0 has two answer candidates, BERT receives two inputs: one input containing ANSWER CANDIDATE A and a second input containing ANSWER CANDIDATE B. Each input has the question and the answer candidate's tokens in segment one, and the passage tokens in segment two. As in Chapter 3, we separate the question's and answer candidate's tokens in segment one using a special separator token.

II. **BERT + question-answer**: BERT fine-tuned on our MCScript2.0 subset. For each question, the model queries an index using the QUESTION-ANSWER strategy described

in Section 4.2.1. The retrieved information is concatenated at the beginning of the text passage. We then build the inputs in the same way we do for the baseline model.

III. **BERT + question-passage-answer**: BERT fine-tuned on our MCScript2.0 subset. For each question, the model queries an index using the QUESTION-ANSWER strategy described in Section 4.2.2. The retrieved information is concatenated at the beginning of the text passage. We then build the inputs in the same way we do for the baseline model.

4.4 Results

Table 4.2 summarizes the results of the three models using both of our retrieval strategies and querying both Visual Genome's and ConceptNet's indexes. Unlike the results of our preliminary experiment in Chapter 3, using image descriptions from Visual Genome showed little improvement on the dev set over the results of the baseline model and there was a decrease on the performance on commonsense questions in the test set. On the other hand, when we retrieve information from ConceptNet the model does improve its performance as shown in both dev and test sets.

Index	Visual Genome		ConceptNet	
Model	Dev	Test	Dev	Test
BERT	0.730	0.728	0.730	0.728
BERT + QUESTION-ANSWER	**0.735**	0.708	**0.737**	**0.735**
BERT + QUESTION-PASSAGE-ANSWER	**0.733**	0.723	0.729	**0.740**

Table 4.2 Accuracy on commonsense questions from MCScript2.0. The results are the average of three runs using different random seeds.

Intuitively, the QUESTION-ANSWER and QUESTION-PASSAGE-ANSWER strategies should retrieve a different set of results. The latter takes into account the three core elements of reading comprehension, so it should contain more meaningful information that the former. However, our results suggest that this was not the case. Both retrieval strategies had the same performance on the dev set, when we query the Visual Genome index, and the QUESTION-ANSWER strategy seem to be better when querying ConceptNet. Only the results on the test set when using ConceptNet showed an improvement of the QUESTION-PASSAGE-ANSWER strategy over QUESTION-ANSWER.

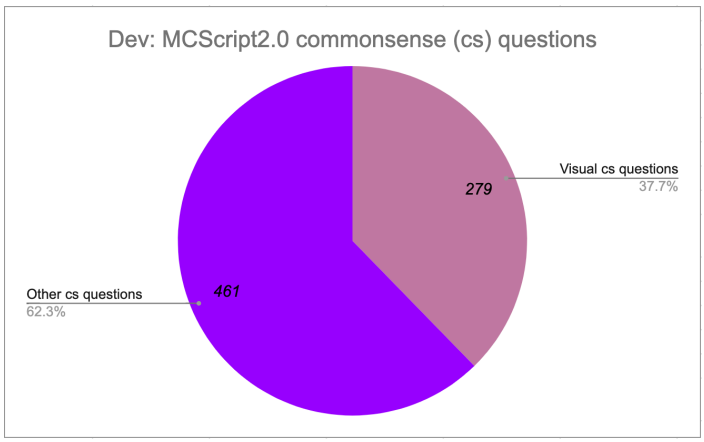

Fig. 4.3 Distribution of VISUAL COMMONSENSE QUESTIONS in our dev set with 740 commonsense questions total.

The two models that query an index for external information include the retrieved results in the same way: they both concatenate that information to the passage. Thus, the difference in the accuracy on commonsense questions can only come from what it is contained in those retrieved results. In the next section, we take a closer look at what is being retrieved from the indexes.

4.5 Evaluation

Accuracy is a metric that tell us how many times the predictions of a model are correct. We are assuming that whenever the model is not able to choose the correct answer candidate, it is because the model lacks knowledge about what is being asked. By querying an index to retrieve information, we expected to feed the model with more knowledge, which would lead to an increase on its accuracy. If the accuracy did not increase, it does not necessarily mean that the retrieved information had no *meaningful* information. There are two possible scenarios: (1) the retrieved result set does not contain useful commonsense knowledge or (2) the retrieved result set contains useful commonsense knowledge, but the model is not being able to process it. We perform our error analysis considering both perspectives.

4.5.1 Retrieval inspection

Similar to how we did in Chapter 3, we selected a subset of commonsense questions from the 740 available in our dev set. These questions are those for which the required knowledge can be learned visually. We refer to these questions as VISUAL COMMONSENSE QUESTIONS. Figure 4.3 shows the proportion of the VISUAL COMMONSENSE QUESTIONS found in the dev set.

For each question, we manually inspected the set of top 50 results returned by the QUESTION-ANSWER and QUESTION-PASSAGE-ANSWER strategies. Recall from Section 4.2 that in either case we remove duplicates and results with less than three tokens to avoid image descriptions like *a blue blanket* that are not so informative. We focused on the set of top 50 results rather than in the set with the top 10 because to verify if either Visual Genome or ConceptNet do have meaningful commonsense knowledge. By *meaningful* we refer to an image description or a surface text that contains the *piece of knowledge* required to answer a question. Consider the two examples below:

(3) **Where** *was the hot dog placed?*

(4) **What** *were the ingredients needed for?*

Question 3 requires knowledge about the common place of a hot dog and question 4 requires knowledge about an activity in which we would use *ingredients*. A *meaningful* image description or surface text would contain such a knowledge.

Our manual inspection confirmed that there was at least one image description with meaningful information for 56.99% of the questions (i.e., 159 questions). As for ConceptNet, we were able to find a meaningful surface text 52.33% of the times. Table 4.3 shows some examples of the meaningful information found. Most of the times we found a meaningful image description in Visual Genome, we also found a meaningful surface text in ConceptNet. Out of the 279 analyzed questions, there were 53 for which there was a meaningful image description but there was not a meaningful surface text. Similarly, there were 40 questions for which was a meaningful surface text but there was not a meaningful image description. Image descriptions tend to be shorter than surface text, but as shown in Table 4.3, the information they provide is essentially the same.

4.5.2 Model inspection

As previously mentioned in Section 3.4.2, BERT builds two inputs for each question and computes a softmax over the two choices to predict which one contains the correct answer candidate. In other words, BERT chooses the input with the highest logit value. In terms of accuracy, if the input with the ANSWER CANDIDATE A has a logit value of 51% and the input with the ANSWER CANDIDATE B has a logit value of 49%, BERT will predict the former to be the correct answer candidate. The same thing would happen if the logit values were 82% and 18%, respectively. Assuming ANSWER CANDIDATE A is the correct answer candidate in

Question	Answer candidate A	Answer candidate B	Image description	Surface text
What did they turn on?	kettle	stove	*two devices turned on*	*switch is related to turn on*
What was filled with water?	The pot	sink	*a container filled with water*	*something you find in a container is water*
What was used to dry off?	a towel	clothes	*towel to use for drying off*	*towel is for drying off*
What did they hold up that they bought from the bar?	Their food	Their drinks	*liquors available at the bar*	*something that might happen when you hang out at the bar is buy a drink*
What was built with the larger logs?	a cabin	the campfire	*wooden logs for the fire*	*wood is used in making a campfire*
What did mom's gift do nicely in a box?	opened	fit	*box that a gift would come in*	*a box is for wrapping for a present*
What was occuring outside?	Cooking	A rainbow	*scene happening outside here*	*view is related to outside*
What did they take out of the plastic and start to break up with their hands?	bread	koi fish	*hands breaking apart croissant*	-
What did they walk to the car with?	Two new t-shirts	shopping bag	-	*something that might happen while buying products is carrying them to the car*

Table 4.3 Sample of meaningful image descriptions and surface texts. The correct answer candidate is underlined. A dash line (-) indicates there was no meaningful entry found in the top 50 results.

this example, the prediction would be correct. The difference is the *confidence* of the model; clearly, in the second case the model is more confident about ANSWER CANDIDATE A being the correct one.

It is possible that the logit value of the CORRECT ANSWER CANDIDATE increases when the model queries an index for external information, but the logit value is still smaller than the logit of the INCORRECT ANSWER CANDIDATE. In that case, we would not see an impact on

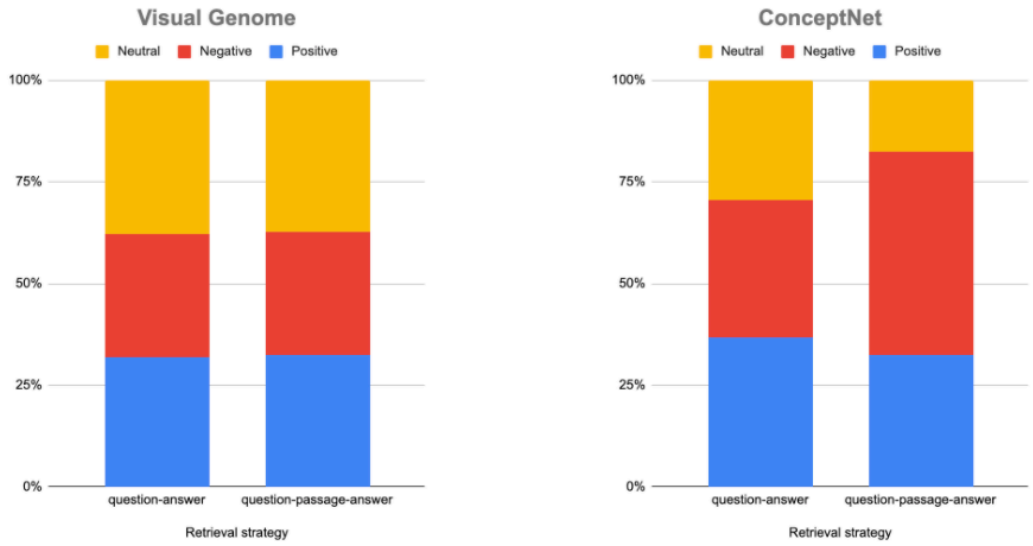

Fig. 4.4 Overall impact on the 740 commonsense questions from the dev set. The impact was positive, negative or neutral depending on the change on the logit value of the CORRECT ANSWER CANDIDATE.

accuracy since the model will still fail to predict the correct one. We need to take a closer look at the logit values of each input to measure the impact of using external information. That impact could be classified as one of the following:

I. **Positive impact**: If there is an **increase** on the logit value of the the CORRECT ANSWER CANDIDATE.

II. **Negative impact**: If there is an **decrease** on the logit value of the the CORRECT ANSWER CANDIDATE.

III. **Neutral impact**: If there is **no change** on the logit value of the the CORRECT ANSWER CANDIDATE.

Figure 4.4 illustrates the impact on the 740 commonsense questions in the dev set. When the model queries Visual Genome for external information, the retrieved information has a positive impact on approximately 30% of the questions. The number of times the additional information harms the performance of the model has a similar rate of 30%. The remaining percentage shows that the retrieved information had no impact at all. It is interesting to see that the impact on BERT's predictions in terms of the logit values did not change, whether we queried the index using the QUESTION-ANSWER strategy or QUESTION-PASSAGE-ANSWER.

The behaviour was different when the model queries ConceptNet instead. There is a clear advantage of using the QUESTION-ANSWER strategy, which results on a positive impact

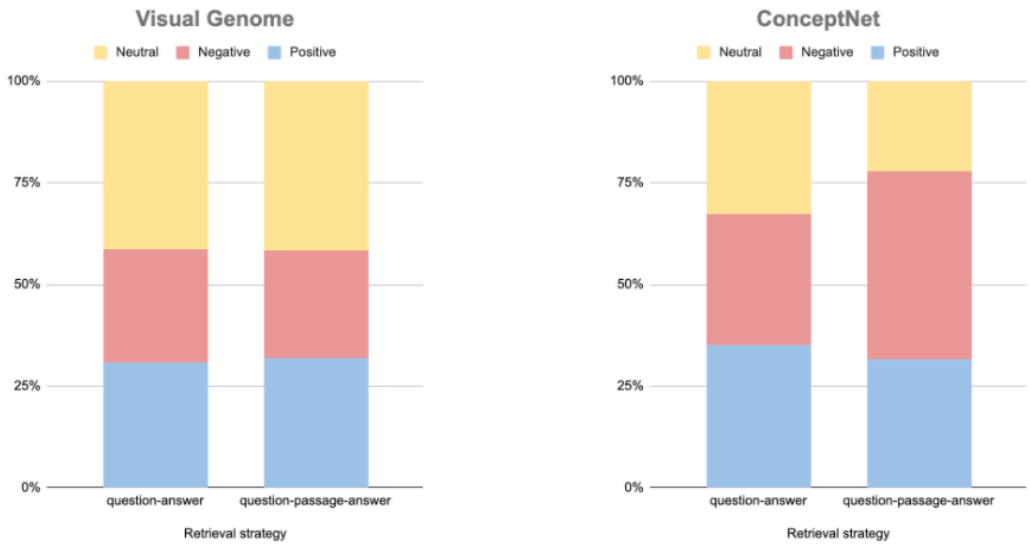

Fig. 4.5 Overall impact on the 279 VISUAL COMMONSENSE QUESTIONS from the dev set. The impact was positive, negative or neutral depending on the change on the logit value of the CORRECT ANSWER CANDIDATE

on around 37% of the questions. This was higher than the positive impact achieved when using either strategy on the Visual Genome index. The positive impact achieved by using the QUESTION-PASSAGE-ANSWER strategy is similar to the one observed when the model queries Visual Genome. However, querying ConceptNet in this way harms the predictions of the model in a bigger rate. Figure 4.5 shows the behaviour when we measure the impact on the 279 VISUAL COMMONSENSE QUESTIONS. It is almost identical to the behaviour observed in the full set of commonsense questions. This raises the question of the model's ability to take advantage of the additional information. Our manual retrieval inspection in Section 4.5.1 showed that there is at least one meaningful entry from both indexes for VISUAL COMMONSENSE QUESTIONS. We further explore the model's performance with an additional experiment.

5-fold cross-validation

We test how the model makes use of the additional information. On our manual inspection in Section 4.5.1 we identified 159 questions for which there was a meaningful image description in Visual Genome and 146 questions for which there was a meaningful surface text from ConceptNet. Our intuition is that if we append that meaningful information to the input with the correct answer candidate, the overall performance of the model will improve. For this experiment, we merged the 12,732 questions in our latest TRAIN set and the 279 VISUAL COMMONSENSE QUESTIONS from our DEV set. Once merged, we split the data into 5

Fold	BERT	BERT+Visual Genome	BERT+ConceptNet
Fold 1	0.729	0.749	0.740
Fold 2	0.722	0.706	0.720
Fold 3	0.690	0.714	0.701
Fold 4	0.740	0.734	0.738
Fold 5	0.724	0.737	0.712
Average	0.721	0.728	0.722

Table 4.4 5-fold cross-validation accuracy. After the + sign, the name of the index that was queried.

parts, making sure that the 279 VISUAL COMMONSENSE QUESTIONS are together in one part. Under this experimental setting, each input has only one image description or surface text, depending on the index used. More specifically, we add information in either of two ways:

I. If the question belongs to the set of questions for which we found a meaningful image description/surface text, we append the meaningful image description/surface text to the input with the CORRECT ANSWER CANDIDATE. For the input with the INCORRECT ANSWER CANDIDATE we add the top ranked result retrieved using the QUESTION-ANSWER strategy.

II. For any other of the questions, we query an index using QUESTION-ANSWER retrieval and add, to both inputs, the top ranked result.

Table 4.4 details the results. The small improvement over the BERT's baseline accuracy suggests that adding information to the input has little impact on the model's predictions, regardless of whether the extra information has something meaningful or not.

4.6 Conclusion

In this chapter, we aimed to automate the retrieval process introduced in Chapter 3 and replicate the improvement observed over the baseline model evaluating on a bigger subset. Also, we indexed the information in a well-known commonsense knowledge base to compare its content with that of the image descriptions. We proposed two retrieval strategies that are independent of the index we chose to query. A manual inspection of the retrieved results confirmed that both strategies retrieve information relevant to a given passage, question and an answer candidate. However, our model was not able to make use the retrieved information

to improve its performance on commonsense questions. How information flows internally in BERT is a recent active area of research. We leave to future work to focus on BERT's internal architecture.

Chapter 5

Conclusion

In this thesis, we explore the current state of NLP systems in terms of human-level language understanding. As these systems continue to evolve, there are more questions on their real capabilities. Knowledge is a core element of linguistic capability and the knowledge human's have build through their life is very broad. This work focused on an specific type of knowledge whose main characteristic is that it is difficult to explain: tacit knowledge. In summary, the contributions of this work are as follows:

I. We targeted temporal relation extraction, one of the most difficult tasks in NLP, and identified two core challenges. Our proposed adapted system achieves state-of-the-art performance on this task, but it still shows a gap with respect to human performance. Our main finding is the identification of a high incidence of nouns as events. We showed that increasing the training data does not help the model grab the concept of an event's *duration*. The main problem with having nouns as events is that the Linguistic means to handle it, *aspect*, cannot be applied to them since aspect is a property of verbs.

II. The challenge identified on a temporal knowledge-based task could potentially be alleviated by relying on an external knowledge base. We explored this possibility on a task whose required knowledge is closely related to the nature of temporal knowledge: commonsense knowledge. We proposed image descriptions as a candidate source for commonsense knowledge acquisition. Our results on a preliminary experiment showed that a state-of-the-art model evaluated on a machine reading comprehension task can benefit from it.

III. We extended the above exploration, designing two retrieval strategies and testing with an additional source to image descriptions: a commonsense knowledge base. These

two text sources differ in how they were created, but we identified that they both contain similar commonsense information.

Tacit knowledge is build upon our experiences, which represents a limitation to a computer system. Nevertheless, our exploration study shows that there is a fair amount of this knowledge in data. However, we have yet to improve how are we representing this knowledge. Our experiments presented in Chapter 3 and Chapter 4 follow the standard approach of adding knowledge into BERT's input as done in previous work like Jain and Singh (2019) and recent work by Petroni et al. (2020). Our immediate future work is to use embeddings similar to TriAN (Wang et al., 2018), the top public model on the MCScript (Ostermann et al., 2018) shared task.

Our motivation to rely on image descriptions is based on the intrinsic relationship between knowledge acquisition and human experience. Even though image descriptions are generated from a visual input, in the end, they are just text. Our next step includes a *multi-modal* approach where we will use an embedding representation of the image from where the chosen image description comes from. This proposal is close in spirit to the multi-modal approach of Ororbia et al. (2019), where a model is trained jointly on corresponding linguistic and visual data. For instance, for the question *What was used to dry off?* we retrieved the description *towel to use for drying off* and appended it to BERT's input. In future work, we will use an embedding representation of the image description and an embedding representation of the towel image from the description refers to.

Our experiments rely on BERT's transformer self-attention mechanism to implicitly model the relation between an input question, passage and answer candidate. The results of our last experiment in Chapter 4 suggest that the model is not able to attend to additional text that contains meaningful information to relate a question with an answer candidate. It was out of the scope of this thesis to analyze how the attention scores are propagated, but it is an important step towards language understanding. Thus, a second step in our future work is to design a probe to pinpoint why meaningful information does not have the expected positive impact on the correct answer candidate's prediction.

References

Allen, J. F. (1990). Readings in qualitative reasoning about physical systems. chapter Maintaining Knowledge About Temporal Intervals, pages 361–372. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Andersen, R. W. (1990). Unpublished lecture in the seminar on the acquisition of tense and aspect.

Bethard, S., Savova, G., Palmer, M., and Pustejovsky, J. (2017). Semeval-2017 task 12: Clinical tempeval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572.

Bybee, J. L., Perkins, R. D., and Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world*, volume 196. University of Chicago Press Chicago.

Chikka, V. R. (2016). Cde-iiith at semeval-2016 task 12: Extraction of temporal information from clinical documents using machine learning techniques. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1237–1240.

Collins, H. (2010). *Tacit and explicit knowledge*. University of Chicago Press.

Costa, F. and Branco, A. (2012). Aspectual type and temporal relation classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 266–275. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dligach, D., Miller, T., Lin, C., Bethard, S., and Savova, G. (2017). Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 746–751.

Forbes, M., Holtzman, A., and Choi, Y. (2019). Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Fries, J. A. (2016). Brundlefly at semeval-2016 task 12: Recurrent neural networks vs. joint inference for clinical temporal information extraction.

Gordon, J. and Van Durme, B. (2013). Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2009). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99.

Huang, L., Bras, R. L., Bhagavatula, C., and Choi, Y. (2019). Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Jain, Y. and Singh, C. (2019). Karna at coin shared task 1: Bidirectional encoder representations from transformers with relational knowledge for machine comprehension with common sense. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 75–79.

Johnson, J., Douze, M., and Jégou, H. (2017). Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Jones, X. (2015a). Semeval-2015 task 6: Clinical tempeval. In for Computational Linguistics, A., editor, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015): 4-5 June 1996; Baltimore*, pages 806–814. Stoneham: Butterworth-Heinemann.

Jones, X. (2015b). Semeval-2016 task 12: Clinical tempeval. In for Computational Linguistics, A., editor, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2015): 4-5 June 1996; Baltimore*, pages 806–814. Stoneham: Butterworth-Heinemann.

Jung, H., Allen, J., Blaylock, N., de Beaumont, W., Galescu, L., and Swift, M. (2011). Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 workshop*, pages 146–154.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kintsch, W. and Walter Kintsch, C. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.

Klein, W. (2013). *Time in language*. Routledge.

Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. (2018). The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Krohn, J., Beyleveld, G., and Bassens, A. (2019). *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. Addison-Wesley Professional.

Kulyk, V. (2006). Constructing common sense: Language and ethnicity in ukrainian public discourse. *Ethnic and racial studies*, 29(2):281–314.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Lee, H.-J., Xu, H., Wang, J., Zhang, Y., Moon, S., Xu, J., and Wu, Y. (2016). Uthealth at semeval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1292–1297.

Leeuwenberg, A. and Moens, M.-F. (2016). Kuleuven-liir at semeval 2016 task 12: Detecting narrative containment in clinical records. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1280–1285.

Leeuwenberg, A. and Moens, M.-F. (2017). Structured learning for temporal relation extraction from clinical records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Li, P. and Huang, H. (2016). Uta dlnlp at semeval-2016 task 12: deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1268–1273.

Li, P., Yasuhiro Shirai, P., and Shirai, Y. (2000). *The Acquisition of Lexical and Grammatical Aspect*. Studies on language acquisition. Mouton de Gruyter, Berlin, New York.

Lin, C., Miller, T., Dligach, D., Bethard, S., and Savova, G. (2016). Improving temporal relation extraction with training instance augmentation. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 108–113.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Liu, H. and Singh, P. (2004). Conceptnet: a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*.

Minsky, M. (2007). *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster.

Miwa, M. and Bansal, M. (2016). End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116.

Mukuze, N., Rohrbach, A., Demberg, V., and Schiele, B. (2018). A vision-grounded dataset for predicting typical locations for verbs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ororbia, A., Mali, A., Kelly, M., and Reitter, D. (2019). Like a baby: Visually situated neural language acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5127–5136, Florence, Italy. Association for Computational Linguistics.

Ostermann, S., Modi, A., Roth, M., Thater, S., and Pinkal, M. (2018). Mcscript: A novel dataset for assessing machine comprehension using script knowledge. *arXiv preprint arXiv:1803.05223*.

Ostermann, S., Roth, M., and Pinkal, M. (2019). Mcscript2. 0: A machine comprehension corpus focused on script events and participants. *arXiv preprint arXiv:1905.09531*.

Ovchinnikova, E. (2012). *Integration of world knowledge for natural language understanding*, volume 3. Springer Science & Business Media.

Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A. H., and Riedel, S. (2020). How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019). Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., and Mani, I. (2005). The specification language timeml. *The language of time: A reader*, pages 545–557.

Pustejovsky, J. and Stubbs, A. (2011). Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rashkin, H., Sap, M., Allaway, E., Smith, N. A., and Choi, Y. (2018). Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sadeghi, F., Kumar Divvala, S. K., and Farhadi, A. (2015). Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011). Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952.

Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y. (2019). Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Schank, R. C. and Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Smith, C. S. (1983). A theory of aspectual choice. *Language*, pages 479–501.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., et al. (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2018). Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

UzZaman, N., Llorens, H., and Allen, J. (2012). Evaluating temporal information understanding with temporal question answering. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 79–82. IEEE.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 1–9.

Vendler, Z. (1957). Verbs and times. *The philosophical review*, 66(2):143–160.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 75–80.

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62.

Wang, L., Sun, M., Zhao, W., Shen, K., and Liu, J. (2018). Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. *arXiv preprint arXiv:1803.00191*.

Werbos, P. J. et al. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Xu, K., Feng, Y., Huang, S., and Zhao, D. (2015). Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv preprint arXiv:1506.07650*.

Yatskar, M., Ordonez, V., and Farhadi, A. (2016). Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

Zwaan, R. A. (1999). Situation models: The mental leap into imagined worlds. *Current directions in psychological science*, 8(1):15–18.

Appendix A

Miwa and Bansal's model adaptation

A.1 Settings

A.1.1 Sentence-level annotations

We used the Clinical Language Annotation, Modeling and Processing (CLAMP) toolkit[1] for tokenization and sentence boundary detection. We then matched all entities' spans from the gold standard with the sentence offsets on the CLAMP output to identify those within the same sentence. Therefore, we created new annotations containing a pair of words, their offsets in the sentence, the temporal relation between them marked on the gold standard, and the directionality of the arguments. Example 5 shows an annotation of the TLINK—CONTAINS*(lifelong, nonsmoker)* in the sentence *"He is a lifelong nonsmoker."* Note that no entity type (i.e., EVENT or TIMEX3) or any of its associated attributes are included.

(5)

T1	Term 8 16	lifelong
T2	Term 17 26	nonsmoker
R1	ContainsSource-ContainsTarget	Arg1:T1 Arg2:T2

The THYME corpus does not identify instances where two entities have none of the TLINK relations. Hence, we define a NONE label and apply it as follows: since any two EVENT/TIMEX3 can be a candidate pair, we take all entities in a sentence to generate all pair permutations as candidates. Pairs that do not have any temporal relation are then labeled as NONE. Therefore, for entities e_1, e_2, and e_3 in a sentence where e_1 CONTAINS e_2, pair (e_1, e_2) is considered as a positive instance while the resulting candidate pairs from our procedure (e_1, e_3), (e_2, e_1),

[1] http://clinicalnlptool.com/index.php

(e_2, e_3), (e_3, e_1), and (e_3, e_2) are considered as negative instances. Due to the large number of negative instances produced (1:3 ratio of positive to negative examples), for a sentence with entities e_4, e_5, and e_6 with a TLINK: e_5 BEFORE e_6—(or any TLINK but CONTAINS)—no negative instances were generated.

A.1.2 Implementation and training

This study used Miwa and Bansal's 2016 implementation, available at https://github.com/tticoin/LSTM-ER. It followed their training settings, updating the model parameters (including weights, biases, and embeddings) by backpropagation through time (Werbos et al., 1990) and Adam (Kingma and Ba, 2014) with gradient clipping, parameter averaging, and L2-regularization. Dropout (Srivastava et al., 2014) was applied to the embedding layer and to the final hidden layers for relation classification. The hyper-parameters used were their default hyper-parameters for the SemEval-2010 Task 8: initial learning rate (1e-6), regularization parameter (1e-6), input dropout probability (0.5), output dropout probability (0.3), gradient clipping size (1), and number of epochs (63).

A.2 Multi-classification performance

A.2.1 In-domain word embeddings

Once we verified the adopted model gave competitive results on the narrative container identification task, we focused on increasing the system's recall. We, therefore changed the default word representations trained on Wikipedia for in-domain word embeddings. Word representation depends on the words in context, and since the clinical domain is a specific field with different vocabulary from those used in the general domain, we expected the model to benefit from a resource like PubMed. However, our results suggest that this does not have a significant impact on most TLINKS. Only BEGINS-ON and ENDS-ON recall considerably improved. This limited improvement can be attributed to the data size. The subset of PubMed abstracts used to train our in-domain word embeddings is smaller than the Wikipedia dump on which Miwa and Bansal's 2016 default word embeddings were trained. A possible method of improving word representation in a temporal-aware context is to rely on transfer learning. For this purpose, pre-training BERT (Devlin et al., 2018) on general domain temporal data and fine-tuning on the Clinical TempEval task could lead to interesting results.

A.2.2 Down-sampling negative examples

We still witnessed an imbalance between precision and recall despite the fact that we increased recall by using in-domain word embeddings. Moreover, our results are still below UTHealth's recall score (highest on CONTAINS identification task). By filtering EVENT pairs as described in Section 2.2.2 experimental setting (IV), the NONE class was reduced by 10%. This further improved the recall for most TLINKS except for ENDS-ON. A McNemar's test on the results of *PubMed word emb* and *PubMed word emb + FNE* yields a p-value of 0.006, which is significant at the 0.05 level. This means that by filtering negative examples, the model's proportion of errors decreases with respect to *PubMed word emb*.