A closer look at learning relations from text

Katrenko, S.

Citation for published version (APA):
1 INTRODUCTION

1.1 SETTING THE SCENE

Understanding of human intelligence is not possible without understanding the language humans use. Even though daily communication is seemingly easy for humans, they need to accomplish many tasks by resolving ambiguities, integrating information and inferring new knowledge. Moreover, machine intelligence is often defined in relation to human intelligence. Recall the definition of machine intelligence in the famous Turing test [190] which is set as follows. There are three rooms in one of which is located a human interrogator, and in the other two a human and a computer. The human interrogator is allowed to put questions but he is not aware of who he is having a conversation with. If he fails to distinguish between a human and a machine, we can conclude that the machine is intelligent. In the ‘imitation game’ proposed by Turing the initial question “Can machines think?” is rephrased by whether machines can act as humans. In other words, one is no longer interested in how a machine arrived at its action as long as this action resembles the one of human.

If many centuries ago there was a strong oral tradition where folklore stories and songs were told to the next generations in the spoken word, in the modern world a lot of information is conveyed in the written form. With the abundance of information that can be found in various sources such as texts, databases, ontologies and others, it became important to be able to process it automatically. One of the most challenging tasks is understanding of natural language texts. In a nutshell, natural language understanding (NLU) is a complex process that requires solving various puzzles and putting them together into a coherent picture. NLU is often seen as

a process of hypothesis management in which the linguistic input is sequentially scanned as the system considers alternative interpretations (Jurafsky and Martin [78])

NLU can be decomposed in several modules one of which could be understanding of relations. Humans use relations in their everyday life either to describe some new things by referring to what is already known or to compare different concepts. Relations can be introduced at different stages, from relations between textual fragments (e.g., discourse relations) to relations between word senses. Consider, for instance, an example [1.1] from Lascarides and Asher [95].
(1.1) (a) The council built the bridge.
    (b) The architect drew up the plans.

Here, by knowing that the event of drawing a plan usually precedes an event of constructing something, we may infer that the event in (1.1b) contributes to the event in (1.1a) and the discourse relation between the two is elaboration. If we however raise the question of who drew up the plans, then we are looking for an instance of the Product - Producer relation which is defined on the more fine-grained level of word senses. In other words, given the type of relation and one of its arguments (‘Product’) we should be able to fill in the other argument (‘Producer’) in such a way that the relation holds.

So far, many linguistic theories \[95\] have been proposed to account for discourse relations. They introduced various types of rhetorical relations such as elaboration (shown above), explanation, result and others. Identifying discourse relations automatically is a difficult task but it is essential for some other areas such as dialogue management. Computational models of recognizing this type of relations make use of linguistic theories and often rely on discourse connectives. The other type of relations that we mentioned above is relations between concepts, which are often referred to as lexico-semantic or semantic relations. The large body of research on semantic relations focused on their representation, properties \[34, 132\], perception by humans and most recently, automatic recognition \[2, 15\].

Despite the fact that a lot of work has been done on automatic relation discovery in the past few decades, it remains a popular research topic. The main reason for the keen interest in relation recognition lies in its utility. Once semantic relations are identified, they can be used for a variety of applications such as question answering, ontology construction, hypothesis generation and others. An example which fits a typical question answering scenario was given above. Given (1.1a) and (1.1b), the correct answer to “Who drew up the plans?” should be ‘the architect’ but not ‘the council’. For ontology learning, it is usually necessary not only to recognize instances of the existing concepts but also establish relationships between them. The most common relationship here is hypernymy, where all concepts are taxonomically organized \[179\]. Yet another application of relations is constructing a new hypothesis given the evidence found in text \[183\]. This type of knowledge discovery is often based on co-occurrence analysis. Imagine there are two articles both of which study some property \(A\). In addition, each of them describes a substance that has this property \(A\). If the substances in question are different, we may expect that the fact they both share \(A\) can lead to a new relationship between the substances. It has been shown that scientific discovery from text in many cases was corroborated via experiments in the laboratories.
Another reason why extraction of semantic relations is still of interest lies in the diversity of relations. On the very general level, semantic relations fall into generic and domain-dependent categories. Many existing information extraction systems were originally designed to work for generic data [61] but it became evident that they should be adapted if one intends to use them in a new domain. To be able to produce similar results on domain-dependent data, a given system should be able to handle terminology. For instance, relation extraction in the biomedical domain would require an accurate recognition of named entities such as gene names while in the food informatics field it should be able to account for other types of named entities such as toxic substances.

To summarize, we address in this thesis several aspects of automatic relation recognition, better understanding of which should allow us to extract relations more precisely.

1.2 VL-e Project

The research presented in this thesis has been done within the subproject Adaptive Information Disclosure (AID) which was part of the larger Virtual Lab environment for e-science (VL-e) project. The VL-e project focused on bringing together distributed computing, high performance networking and development of scientific applications. Further, the VL-e programme consisted of four subprogrammes such as P1 (e-Science in applications), P2 (Generic Virtual Laboratory methodology), P3 (Large-scale distributed systems), and P4 (Scaling up to and validating in ‘real-life applications’).

Within VL-e, the AID project was placed into the P2 subprogramme. AID had several goals, one of which was studying natural language processing capabilities and exploring machine learning for information extraction. As it is stated in Hertzberger and Vos [71],

The program will explore new avenues in machine learning for information management by employing NL analysis tools, grammar induction, and related tools in contexts where the domain of application is partially known beforehand. This will involve drawing relevant information from ontologies and making it available for guided induction.

Fig. 1 highlights various areas that have been studied within AID. The work presented here falls into the part named ‘ontology learning’. During the project we examined existing named entity recognition solutions (both supervised and unsupervised) [80, 119, 82, 174] and proposed novel methods for extracting relations from text [79, 83]. We addressed the problem of ontology population and showed that information that is already available in existing ontologies can be used to improve information extraction performance. The information extraction results can
be employed further in several possible ways. For instance, discovered instances of existing concepts can be added to the ontology [191] or they can be integrated with other information sources to guide knowledge discovery [154]. AID aimed at providing generic solutions that would be applicable to a wider range of domains. We have experimented with both domain-dependent and generic data and showed that the methods we introduced are generic and work well for both types of data. Depending on the machine learning method, learning can be a long, time-consuming process. We have studied how to speed it up by exploring the possibilities offered by the distributed ASCI Supercomputer (DAS-3). Our experiments on local alignments kernels discussed in Chapter 6 of this thesis were all conducted on DAS-3.

1.3 Research Questions

Our main goal is providing a computational model for relation extraction. Having this in mind, we consider different types of relations starting from textual entailment up to semantic relations.

The research questions that we formulate in this thesis are as follows.

1. What is the role of syntactic information for relation recognition?

Most current approaches to relation extraction make use of syntactic information. It is recognized that ‘subject-verb-object’ tuples can be
employed for accomplishing this task but semantic relations can be syntactically realized in many more ways. It is also unclear what types of syntactic information are useful. To answer this question, we focus on dependency structures and consider two types of relations, sentential (textual entailment) and binary biomedical relations. The role of syntactic analysis is discussed in Chapter 4 and Chapter 5.

2 How can we effectively incorporate prior knowledge in the learning process?

Intuitively, any additional relevant information should help improve relation learning. For natural language data such information can either be gathered from the large corpora or be collected from the existing structured resources such as ontologies, taxonomies or others. This leads to the question of what are the ways of using such information during the learning process. This question is addressed in Chapter 6 by exploring local alignment kernels for generic and domain-specific relations.

3 Is it possible to learn cognitively plausible semantic constraints (types) for various generic relations? Can they be used for relation recognition?

Provided with a relation, humans can relatively easily give some examples of it. Moreover, they can also describe it, often referring to what types of arguments this relation may have. For instance, if we think of the Part - Whole relation, a possible definition could be “for two entities X and Y, in any situation where X is part of Y, Part - Whole takes place”. In addition, we may provide some examples as <steel, car>, <cotton, jacket> where ‘steel’ is part of ‘car’ and ‘cotton’ is part of ‘jacket’. If we look at these examples more precisely, we may notice that some arguments can be generalized. In our example a possible generalization would be ‘stuff-object’ where any two entities of type ‘stuff’ and ‘object’ form a meronymic relation. We believe that such generalizations (or semantic constraints) can be useful not only for the actual applications but can also give us more insight into the nature of semantic relations. The problem of semantic constraints discovery is tackled in Chapter 7.

In this thesis we have shown that both syntactic and semantic information is crucial for learning relations. The detailed answers to the research questions are given in the Conclusions chapter.
1.4 METHODOLOGY

We address all research questions by formulating relation recognition as a machine learning problem. We follow a data-driven supervised approach which can be summarized as follows. All instances of semantic relations in text that are annotated are considered positive. Negative instances are created by the closed world assumption where any word pair in the training set which is not labeled as a positive instance is taken to be a negative one. Generating negative instances from the training set according to the closed world assumption is a common practice in machine learning. In this thesis, we consider semantic relations within a sentence and for this reason negative instances are also generated within sentential boundaries. The goal then consists of inferring a learning model (a classifier) that, given new, unseen data would be able to correctly identify relations. Throughout the thesis we discuss several learning algorithms (including meta-learning) and provide a motivation for our choice.

Textual entailment differs from binary semantic relations whereby not a pair of words but a pair of text fragments is annotated. To recognize textual entailment, we propose a solution that can be considered as a combination of unsupervised and supervised learning. Here, mining on syntactic structures is used to output the tree that two structures share. Then, the results of mining are used together with the true label to find what amount of syntactic structure should be shared by two text fragments in order for them to be labeled as positive.

Further, we are interested in both feature engineering and exploring novel methods for relation extraction. Feature engineering is done by using syntactic structures and relies either on studying aspects important for effective relation learning or on discovering common frequent substructures. With respect to the machine learning approaches we (1) use existing methods given the features we constructed, (2) propose a novel method to relation recognition based on the local alignment of sequences.

1.5 OUTLINE

This thesis consists of the following chapters.

Chapter 2 This chapter aims at discussing relations from several aspects. Firstly, it gives a definition of a relation and explores linguistic and logical properties of relations. Secondly, it provides a brief overview of the methods that have been used for automatic relation recognition. Thirdly, we look at several applications and discuss the utility of relations in their context.

Chapter 3 This chapter covers general notions used in machine learning and discusses the methods that are later used.
Chapter 4 In this chapter we discuss textual entailment as an inter-sentential relation and review the approaches that have been proposed to entailment recognition. Further, we present our method based on mining maximal embedded trees.

Chapter 5 In this chapter we zoom in to semantic relations and explore the role of the dependency structures for relation extraction. We introduce a level-based representation which can be considered as a feature engineering step for relation recognition. To test our hypothesis, we present the experimental findings on biomedical data.

Chapter 6 provides a motivation for using more information (prior knowledge) for recognizing semantic relations. We examine a local alignment kernel and propose a solution for accurate relation extraction on both generic and domain-specific data.

Chapter 7 discusses semantic constraints that can be imposed on the arguments of some generic relations. We introduce two methods to derive semantic constraints, provide evaluation by humans, and explore usefulness of the generated constraints for learning generic relations.

Chapter 8 concludes this thesis by presenting the answers to the research questions and discussing the main contributions.