Combining Semi-Unsupervised Acquisition of Corpora and Supervised Learning of Textual Entailment Rules

Fabio Massimo Zanzotto
DISP,
University of Rome “Tor Vergata”,
Roma, Italy
zanzotto@info.uniroma2.it

Marco Pennacchiotti
Computerlinguistik,
Universität des Saarlandes,
Saarbrücken, Germany
pennacchiotti@coli.uni-sb.de

Alessandro Moschitti
DISI,
University of Trento,
Povo di Trento, Italy
moschitti@disi.unitn.it

1 Introduction

As many NLP tasks, textual entailment recognition (RTE) requires large semantic knowledge bases. Unsupervised or semi-supervised knowledge learning methods are very attractive ways for acquiring these knowledge bases from raw texts. Many of these approaches use the Distributional Hypothesis. This can be used to detect equivalence or relatedness between words (e.g., dog and cat), word sequences, and generalized word sequences (e.g., “*X owns Y*” and “*Y belongs to X*”). In the case of RTE, this is only a part of the knowledge that is needed.

Determining whether or not “*Kesslers team conducted 60,643 face-to-face interviews with adults in 14 countries*” entails “*Kesslers team interviewed more than 60,000 adults in 14 countries*” requires two different kinds of knowledge:

- the *equivalence* between “*X conducted Y interviews with Z*” and “*X interviewed Y Z*”
- the *implication rule* that says “*X*” → “*more than Y*” if “*X is bigger than Y*”

The first equivalence can be easily learnt using large corpora and distributional approaches. Yet, the implication rule seems to be out of the scope for these methods.

We are extremely interested in exploring methods to extract from corpora the second kind of knowledge, i.e. *implication rules*. The underlying idea we are following is that these rules can be extracted or exploited using supervised machine learning algorithms over automatically acquired positive and negative textual entailment pairs. We are thus exploring two sub-problems:

- Automatic acquisition of entailment corpora: how to automatically acquire corpora of text-hypothesis entailment and non-entailment pairs. In Section 2 we present an innovative method for this task, leveraging the revisions of documents in collaborative writing systems, such as Wikipedia.
- Definition of a suitable feature space: which is the most suitable feature space to encode examples in order to extract first order entailment rules such as the one described above. In Section 3 we present a novel model to encode these rules in kernel-based feature spaces.

2 Semi-automatically acquired corpora

Our main intuition in using Wikipedia to build an entailment corpus is that the wiki framework should provide a natural source of non-artificial examples of true and false entailments, through its revision system. Wikipedia is an open encyclopedia, where every person can behave as an author, inserting new entries or modifying existing ones. We call *original entry* S_1 a piece of text in Wikipedia before it is modified by an author, and *revision* S_2 the modified text. Our hypothesis is that (S_1, S_2) pairs extracted from the Wikipedia database, represent good candidate of both true and false entailment pairs (T, H). From this source, we can extract pairs like (“*In this regard, some have charged the New World Translation Committee (NWTC) with being inconsistent.*”, “*In this regard, some have charged the New World Translation Committee (NWTC) with not be consistent.*”). These pairs are extremely useful in learning first-order entailment recognition rules as the lexical distance between text and hypothesis is small.

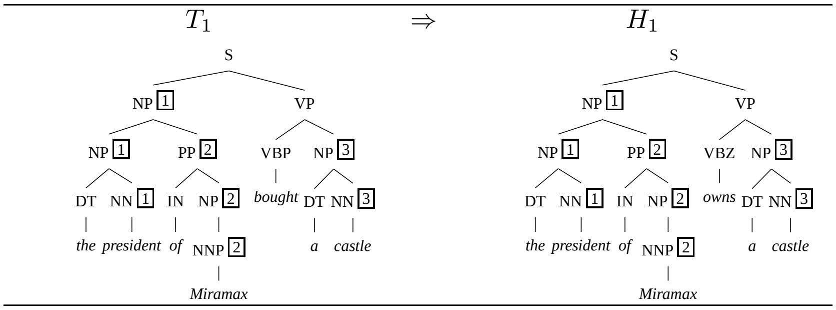


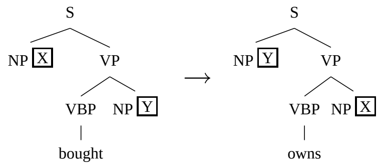
Figure 1: A syntactically analyzed textual entailment pair

3 Learning model

Our first-order entailment rule feature space allows to encode different kinds of rules ranging from lexical-syntactic rules (Zanzotto and Moschitti, 2006), to shallow semantic rules (Pennacchiotti and Zanzotto, 2007) and more complex semantic ones.

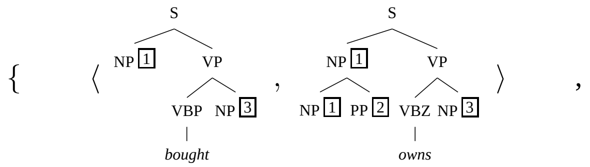
As an example, we here describe a simple first-order syntactic rewrite rule (FOSR) feature space. In this space, each feature f_ρ represents a syntactic first-order or grounded rewrite rule ρ. The rule:

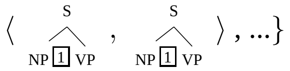
$\rho' = l' \to r'=$



is represented with the feature $< l', r' >$. A (T, H) pair p activates a feature f_ρ if it unifies with the rule ρ. For example, the above feature $f_{\rho'}$ is activated for the example in Fig. 1.

As the full FOSR feature space has an exponential number of features, kernel-based machine learning models such as SVM are usually used to optimize computations. Our kernel is defined as follows. Let $\mathcal{F}(T, H)$ be the set of features that the example (T, H) activates. For example, the set of features $\mathcal{F}(T_1, H_1)$ activated by the example in Fig. 1 is: $\mathcal{F}(T_1, H_1) =$





The kernel function $K((T', H'), (T'', H''))$ that we need to model is then:

$$K((T', H'), (T'', H'')) = |\mathcal{F}(T', H') \cap \mathcal{F}(T'', H'')|$$

The problem of computing this kernel is exponential in the number of variables between T and H (Zanzotto and Moschitti, 2006). We discovered an approximated and efficient version that we proposed in (Moschitti and Zanzotto, 2007).

4 Conclusions

he methods described in the previous sections can be combined together to extract a large set of entailment rules. The application of kernel techniques to very large entailment corpora, harvested automatically from Wikipedia, should indeed guarantee high precision and coverage in the extraction task. At the moment, we are investigating how to best combine this two techniques.

References

Alessandro Moschitti and Fabio Massimo Zanzotto. 2007. Fast and effective kernels for relational learning from texts. In *Proceedings of the International Conference of Machine Learning (ICML)*. Corvallis, Oregon.

Marco Pennacchiotti and Fabio Massimo Zanzotto. 2007. Learning shallow semantic rules for textual entailment. In *Proceedings of RANLP 2007*. Borovets, Bulgaria.

Fabio Massimo Zanzotto and Alessandro Moschitti. 2006. Automatic learning of textual entailments with cross-pair similarities. In *Proceedings of the 21st Coling and 44th ACL*, pages 401–408. Sydney, Australia, July.