Identifying relational concept lexicalisations by using general linguistic knowledge

Maria Teresa Pazienza and **Marco Pennacchiotti** and **Fabio Massimo Zanzotto** [1]

Abstract. This paper analyses how general-purpose semantic hierarchies could be helpful in the construction of one-to-many mappings between the coarse-grained relational concepts and the corresponding linguistic realisations. We propose an original model, the *semantic fingerprint*, for exploiting ambiguous semantic information within the feature vector model.

1 INTRODUCTION

Natural Language Processing (NLP) and Semantic Web (SW) are strictly related disciplines and a beneficial cross-fertilisation is expected. As one of the main problems in the SW is indeed the production of ontologically interpreted web documents, NLP techniques may be fairly useful to bridge the gap between between "ontological relationships" and linguistic forms. This distance depends on the proliferation of linguistic forms denoting ontological relationships. For example, the relationship `teacherOf(Faculty_Member,Course)` is expressed via different realisations such as *Prof. Brown delivers courses on linguistics*, *Prof. Brown teaches courses on linguistics*, or *Brown is the professor of the linguistics course*. This gap has been largely investigated in Information Extraction (IE) [4], where templates are ontological relationships and extraction patterns are linguistic representations of those templates.

In this paper we will address the definition of the one-to-many mappings between coarse-grained relational concepts and the corresponding linguistic expressions. We will evaluate different algorithms and different semantic resources against the specific problem of assigning the correct relational concept given a prototypical linguistic realisation. That is, given what we call a prototypical relational concept form such as "`Faculty_Member` *is the professor of* `Course`" (described as a generalization of a few instances met in the corpus under analysis), identify `teacherOf` as the correct ontological relationship. As we want to investigate the nature of the general semantic knowledge truly required for the task, we propose the notion of *semantic fingerprint* (Sec. 2) to use well assessed machine learning algorithms based on the feature vector model.

2 SEMANTIC FINGERPRINTS FOR LEARNING RELATIONAL CONCEPTS

Some sort of "semantic" generalisation for verbs and nouns in the prototypical relational concept forms may give an important input to cluster these forms in classes. Having, for instance, the form "`Entity` *lose* `Percent`", one of the possibilities to find its equivalence with "share *fall* `Percent`" and distinguishing it from "`Entity` *own* `Percent`" relies on the generalisation of the verb. According to WordNet [2] *lose* and *fall* have two common ancestors (i.e. *change* and *move-displace*), while it does not happen for *fall* and *own*.

The introduction of a conceptual hierarchy in a feature vector is somehow in contrast with the *flatness* of the feature value sets. Thus, these hierarchies should be somehow reduced to a flat set SF where the structure is simply forgot. Words are then mapped to this level of generalisation. Moreover, the unsolved ambiguity in mapping words to senses may create inconsistencies as *certainity* of observations cannot be guaranteed. Features can not have multiple values.

We then propose the notion of *semantic fingerprint* to overcome these problems. A word w (a verb or a noun) will leave its fingerprint $SF(w)$ on the set SF as follows:

$$SF(w) = \{s \in SF | s \text{ generalises } s' \text{ and } s' \in senses(w)\} \quad (1)$$

where $senses(w)$ are all the senses activated by the word w in the considered semantic resource (e.g. WordNet). The machine learning algorithm will select the sense (or the senses) more promising for representing the investigated relationship. The algorithm will therefore also work as task driven sense disambiguator if the semantic information and the way we use it demonstrates to be useful.

Integrating the semantic fingerprint in the feature vector model is straightforward. Given an $S_i = \{true, false\}$ for each element in SF, the subpart of the feature space related to the semantic fingerprint is $S_1 \times ... \times S_n$ where n is the cardinality of SF. Each instance i containing the word w will have the feature value $s_j = true$ if $s_j \in SF(w)$ and $s_j = false$ otherwise.

With the semantic fingerprint abstraction we investigated two "semantic" models against a "bag-of-word" model. These are originated from the assumption that verbs play a relevant role in the problem under analysis. Then, the proposed models are:

$$\text{verb-gen: } V \times W_1 \times ... \times W_n \times VS_1 \times ... \times VS_k \quad (2)$$

$$\text{noun-gen: } V \times W_1 \times ... \times W_n \times NS_1 \times ... \times NS_m \quad (3)$$

where V ranges over all the possible verbs, $W_1 \times ... \times W_n$ represents the "bag-of-word" approach collecting all the verb arguments, $VS_1 \times ... \times VS_k$ is the semantic fingerprint for the verbs, and, finally, $NS_1 \times ... \times NS_m$ is the semantic fingerprint for the nouns. The baseline model, that it is in itself a good model, is called *plain* and it collects verbs and the bag-of-word of the arguments (i.e. $V \times W_1 \times ... \times W_n$).

[1] University of Rome "Tor Vergata", Department of Computer Science, Systems and Production, Roma, Italy email: {pazienza, pennacchiotti, zanzotto}@info.uniroma2.it

3 EXPERIMENTAL ANALYSIS

Experimenting the proposed approach is difficult as large repositories of one-to-many mappings between domain specific ontological relationships and linguistic forms are not easily accessible. We then firstly prepared a test set in order to clarify the final classification task. We produced two different sources of information in order to cross check results. Given a catalogue C of relational concepts, we have produced:

- *classified forms*: a set of one-to-many associations between the concepts in C and the linguistic normalised forms
- *classified sentences*: a set of one-to-many associations between the concepts in C and sentences in the analysed corpus somehow related to the analysed linguistic forms

For the experiments, we used a corpus consisting of financial news (around 12,000 textual news items published from the Financial Times in the period Oct./Dec. 2000). We, firstly, run a *corpus processing phase* selecting around 44,000 forms appearing more that 5 times. Secondly, in the *concept formation phase* a domain expert inspecting the top ranked forms defined 12 target relational concepts. Finally a *classification phase* has been performed by 2 human experts, to which were given two separate set of forms to classify (respectively 3500 and 2200 taken from the first 6500 forms produced in the corpus processing phase). For each form the expert had to decide the correct class. In case of indecision the expert could ask the system to show one or more sentences instance of the form, in order to gain enough information to classify the form itself. Annotators were also asked to classify all the shown sentences. The two data sets, *classified forms* and *classified sentences*, consist, respectively, of the 1091 forms and 6609 sentences. The inter-annotation agreement (computed using 300 forms in common between the two experts and the corresponding 1417 sentences) is 90% on the normalised forms, while the agreement on the sentences is 74%.

Classified forms

Method		*plain*	*verb-gen*	*noun-gen*
Trees	j48.J48	63,91%	63,68%	64,37%
	ID3	59,31%	59,31%	59,54%
	DecStump	26,44%	31,95%	26,44%
Lazy	IB1	58,39%	63,22%	57,70%
	IBk	62,53%	65,98%	60,69%
Rules	j48.PART	59,77%	60,00%	63,22%
Bayes	NaiveBayes	53,33%	58,85%	40,23%
Misc	VFI	59,31%	57,24%	58,39%
	HyperPipes	60,92%	62,76%	62,07%

Classified sentences

Method		*plain*	*verb-gen*	*noun-gen*
Trees	j48.J48	59,19%	64,80%	64,98%
Lazy	IBk	59,19%	54,72%	53,99%
Bayes	NaiveBayes	47,25%	54,03%	42,48%
Misc	VFI	43,81%	52,08%	51,84%
	HyperPipes	31,21%	42,56%	42,48%

Table 1. Results on the sets of *classified forms* and of *classified sentences* (5-fold cross-validation)

The classification problem over the two different proposed data set has been therefore analysed with a pool of algorithms gathered in Weka [7]. This *cross-algorithm validation* can give hints on the relevance and the stability of the chosen feature spaces and on the correctness of the proposed model. For the first set, the *classified forms*, results are reported in Tab. 1. The baseline of the classification is around 27% (naive classification of all the instances in the more probable class). All the algorithms report both in the lexical and the two lexical-semantic spaces better results with respect to the baseline, showing that the chosen features convey the right information for our classification problem. Moreover, the use of the semantic information seems to be relevant, as it emerges in the performance improvement obtained with the majority of the investigated algorithms using the semantic fingerprints on both verbs and nouns; in particular, the verb semantic generalization features seem to be particularly useful. In order to verify how the verb semantic information drives the classification, it can be interesting to examine sample rules produced by a rule based algorithm (j48.PART). For instance the rule ($price = false \wedge job = false \wedge hire = false \wedge succeed = true \wedge entityNE = true \Longrightarrow$ staff movement) indicates that every form containing a verb of *succession* (i.e., a troponym, in the Wordnet sense, of the verb *succeed*) together with an *entityNE* (that is, a company or a person) has to be classified in class staff movement. This semantic generalised rule, according to the Wordnet hierarchy, therefore classifies verbs of *succession* like *enter, supplant, replace, substitute*. Such a general rule can not be captured in a simple lexical space. For the experiment on the *classified sentences* (Tab. 1) we used a reduced pool of algorithm, representative of the different classification methodologies. In this case the baseline is around 40%, corresponding to a naive classification of all the instances in class 5-3. Similarly to the previous experiment, the results show a performance improvement using the verb and noun semantic information.

4 CONCLUSIONS

It is largely agreed that availability of explicit many-to-one mappings between linguistic forms and their corresponding meaning is beneficial for several applications and automatic methods for building these mappings are largely investigated in fields such as Information Extraction [6], Question Answering [5], Terminology Structuring [3], or Paraphrasing [1]. As for all the methods, the use of some previous specific knowledge (not always available) seems mandatory, we tried to attack the problem from a different perspective proposing a method for exploiting well-assessed machine learning algorithm for the problem of learning equivalent surface forms. We obtained some indications that the proposed way to use semantic hierarchies may be helpful in the analysed problem.

REFERENCES

[1] Regina Barzilay and Kathleen McKeown, 'Extracting paraphrases from a parallel corpus', in *Proceedings of the 39th ACL Meeting*, Toulouse, France, (2001).

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

[3] Emmanuel Morin, *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*, Ph.D. dissertation, Univesité de Nantes, Faculté des Sciences et de Techniques, 1999.

[4] Maria Teresa Pazienza, *Information Extraction. A Multidisciplinary Approach to an Emerging Information Technology*, number 1299 in LNAI, Springer-Verlag, Heidelberg, Germany, 1997.

[5] Deepak Ravichandran and Eduard Hovy, 'Learning surface text patterns for a question answering system', in *Proceedings of the 40th ACL Meeting*, Philadelphia, Pennsilvania, (2002).

[6] Ellen Riloff, 'Automatically generating extraction patterns from untagged text', in *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, Portland, Oregon, (1996).

[7] Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, Chicago, IL, 1999.