UNITN: Part-Of-Speech Counting in Relation Extraction

Fabio Celli
University of Trento
Italy
fabio.celli@unitn.it

Abstract

This report describes the UNITN system, a Part-Of-Speech Context Counter, that participated at Semeval 2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals. Given a text annotated with Part-of-Speech, the system outputs a vector representation of a sentence containing 20 features in total. There are three steps in the system's pipeline: first the system produces an estimation of the entities' position in the relation, then an estimation of the semantic relation type by means of decision trees and finally it gives a predicition of semantic relation plus entities' position. The system obtained good results in the estimation of entities' position (F1=98.3%) but a critically poor performance in relation classification (F1=26.6%), indicating that lexical and semantic information is essential in relation extraction. The system can be used as an integration for other systems or for purposes different from relation extraction.

1 Introduction and Background

This technical report describes the UNITN system (a Part-Of-Speech Context Counter) that participated to Semeval 2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals (see Hendrickx *et al.*, 2009). A different version of this system based on Part-Of-Speech counting has been previously used for the automatic annotation of three general and separable semantic relation classes (taxonomy, location, association) obtaining an average F1-measure of 0.789 for english and 0.781 for italian, see Celli 2010 for details. The organizers of Semeval 2010 Task 8 provided ten different semantic relation types in context, namely:

- **Cause-Effect** (CE). An event or object leads to an effect. Example: *Smoking causes cancer.*

- **Instrument-Agency** (IA). An agent uses an instrument. Example: *Laser printer.*

- **Product-Producer** (PP). A producer causes a product to exist. Example: *The growth hormone produced by the pituitary gland.*

- **Content-Container** (CC). An object is physically stored in a delineated area of space, the container. Example: *The boxes contained books.*

- **Entity-Origin** (EO). An entity is coming or is derived from an origin (e.g., position or material). Example: *Letters from foreign countries.*

- **Entity-Destination** (ED). An entity is moving towards a destination. Example: *The boy went to bed.*

- **Component-Whole** (CW). An object is a component of a larger whole. Example: *My apartment has a large kitchen.*

- **Member-Collection** (MC). A member forms a nonfunctional part of a collection. Example: *There are many trees in the forest.*

- **Message-Topic** (CT). An act of communication, whether written or spoken, is about a topic. Example: *The lecture was about semantics.*

- **Other**. The entities are related in a way that do not fall under any of the previous mentioned classes. Example: *Batteries stored in a discharged state are susceptible to freezing.*

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 198–201,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

The task was to predict, given a sentence and two marked-up entities, which one of the relation labels to apply and the position of the entities in the relation (except from "Other"). An example is reported below:

```
``The <e1>bag</e1>
contained <e2>books</e2>,
a cell phone and notepads,
but no explosives.''
Content-Container(e2,e1)
```

The task organizers also provided 8000 sentences for training and 2717 sentences for testing. Part of the task was to discover whether it is better to predict entities' position before semantic relation or viceversa.

In the next section there is a description of the UNITN system, in section 3 are reported the results of the system on the dataset provided for Semeval Task 8, in section 4 there is the discussion, then some conclusions follow in section 5.

2 System Description

UNITN is a Part-Of-Speech Context Counter. Given as input a plain text with Part-Of-Speech and end-of-sentence markers annotated it outputs a numerical feature vector that gives a representation of a sentence. For Part-Of-Speech and end-of-sentence annotation I used Textpro, a tool for NLP that showed state-of-the-art performance for POS tagging (see Pianta *et al.*, 2008). The POS tagset is the one used in the BNC, described at `http://pie.usna.edu/POScodes.html`.

Features in the vector can be tailored for specific tasks, in this case 20 features were used in total. They are:

1. Number of prepositions in sentence.
2. Number of nouns and proper names in sentence.
3. Number of lexical verbs in sentence.
4. Number of "be" verbs in sentence.
5. Number of "have" verbs in sentence.
6. Number of "do" verbs in sentence.
7. Number of modal verbs in sentence.
8. Number of conjunctions in sentence.
9. Number of adjectives in sentence.
10. Number of determiners in sentence.
11. Number of pronouns in sentence.
12. Number of punctuations in sentence.
13. Number of negative particles in sentence.
14. Number of words in the context between the first and the second entity.
15. Number of verbs in the context between the first and the second entity.
16. patterns (from, in, on, by, of, to).
17. POS of entity 1 (noun, adjective, other).
18. POS of entity 2 (noun, adjective, other).
19. Estimate of entities' position in the relation (e1-e2, e2-e1, 00).
20. Estimate of semantic relation (relations described in section 1 above).

Prepositional patterns in feature 16 were chosen for their high cooccurrence frequency with a semantic relation type and their low cooccurrence with the other ones.

The system works in three steps: in the first one features 1-18 are used for predicting feature 19, in the second one features 1-19 are used for predicting feature 20. In the third step, after the application of Hall 1998's attribute selection filter (that evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them) features 12, 14, 16, 19 and 20 are used for the prediction of semantic relation plus entities' position (19 relations in total).

For all the steps I used C4.5 decision trees (see Quinlan 1993) and Cohen 1995's RIPPER algorithm (Repeated Incremental Pruning to Produce Error Reduction). Evaluation for steps 1, 2 and 3 have been run on the training set, with a 10-fold cross-validation, since the test set was relased in a second time. Results of evaluation of step 1, 2 and 3 are reported in table 1 below, chance values (100/number of classes) are taken as baselines, all experiments have been run in Weka (see Witten and Frank, 2005).

I also inverted step 1 and 2 for predicting seman-

Prediction	**Baseline**	**average F1**
step 1	33.33%	98.3%
step 2	10%	29.8%
step 3	5.26%	28.1%

Table 1: Evaluation for steps 1, 2 and 3.

tic relation estimate before entities' position estimate and the average F1-measure is even worse (0.271), demonstrating that entities' position estimate has a positive weight on semantic relation estimate. There are instead some problems with step 2, and I will return on this later in the discussion (section 4).

3 Results

As it was requested by the task, the system has been run 4 times in the testing phase: the first time (r1) using 1000 examples from the training set for building the model, the second time (r2) 2000 examples, the third (r3) 4000 example and the last one (r4) using the entire training set.

The results obtained by UNITN in the competition are not good, overall performance is poor, especially for some relations, in particular Product-Producer and Message-Topic. The best performance is achieved by the Member-Collection relation (47.30%), that changed from 0% in the first run to 42.71% in the second one. Scores are reported, relation by relation, in table 2 below, the discussion follows in section 4.

Rel	**F1** (r1)	**F1** (r2)	**F1** (r3)	**F1** (r4)
CE	23.08%	17.24%	22.37%	26.86%
CW	13.64%	0.00%	13.85%	25.23%
CC	26.43%	25.36%	26.72%	28.39%
ED	37.26%	37.25%	46.27%	46.35%
EO	36.60%	36.49%	37.61%	41.79%
IA	10.68%	7.95%	5.59%	17.32%
MC	0.00%	42.71%	43.08%	47.30%
CT	1.48%	0.00%	4.93%	6.81%
PP	0.00%	0.00%	1.67%	0.00%
Other	27.14%	26.15%	25.80%	20.64%
avg*	16.57%	18.56%	22.45%	26.67%

Table 2: Results. *Macro average excuding "Other".

4 Discussion

On the one hand the POSCo system showed an high performance in step 1 (entities' position detection), indicating that the numerical sentence representation obtained by means of Part-Of-Speech can be a good way for extracting syntactic information.

On the other hand the POSCo system proved not to be good for the classification of semantic relations. This clearly indicates that lexical and semantic information is essential in relation extraction. This fact is highlighted also by the attribute selection filter algorithm that choosed, among others, feature 16 (prepositional patterns), which was the only attribute providing lexical information in the system.

It is interesting to note that it chose feature 12 (punctuation) and 14 (number of words in the context between the first and the second entity). Punctuation can be used to provide, to a certain level, information about how much the sentence is complex (the higher the number of the punctuation, the higher the subordinated phrases), while feature 14 provides information about the distance between the related entities and this could be useful for the classification between presence or absence of a semantic relation (the longer the distance, the lower the probability to have a relation between entities) but it is useless for a multi-way classification with many semantic relations, like in this case.

5 Conclusions

In this report we have seen that Part-Of-Speech Counting does not yield good performances in relation extraction. Despite this it provides some information about the complexity of the sentence and this can be useful for predicting the position of the entities in the relation. The results confirm the fact that lexical and semantic information is essential in relation extraction, but also that there are some useful non-lexical features, like the complexity of the sentence and the distance between the first and the second related entities, that can be used as a complement for systems based on lexical and semantic resources.

References

Fabio Celli. 2010. Automatic Semantic Relation Annotation for Italian and English. (technical report available at `http://clic.cimec.unitn.it/fabio`).

William W. Cohen. 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*. Lake Tahoe, CA.

Mark A. Hall. 1998. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. Technical report available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.148.6025&rep=rep1&type=pdf`.

Iris Hendrickx and Su Nam Kim and Zornitsa Kozareva and Preslav Nakov and Diarmuid Ó Séaghdha and Sebastian Padó and Marco Pennacchiotti and Lorenza Romano and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals. In *Proceedings of the 5th SIGLEX Workshop on Semantic Evaluation*, Uppsala, Sweden.

Emanuele Pianta and Christian Girardi and Roberto Zanoli. 2008. The TextPro tool suite. *In Proceedings of LREC*, Marrakech, Morocco.

John Ross Quinlan. 1993. C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*, San Mateo, CA.

Ian H. Witten and Eibe Frank. 2005. *Data Mining. Practical Machine Learning Tools and Techniques with Java implementations*. Morgan and Kaufman, San Francisco, CA.