Improving Relation Extraction with Anaphora in Italian

Fabio Celli and Massimo Poesio

CLIC-CIMeC, University of Trento,
Italy

Abstract. Semantic Relation Extraction is the task of retrieving information about entities and Anaphora Resolution is the task of linking expressions referring to entities in text. In this paper we use Anaphora Resolution to increase the number of information about entities in order to improve the performance of Semantic Relation Extraction in Italian. We developed three relation extraction systems, based on three different extraction techniques, and we run our systems on the dataset, first without using the Anaphora Resolution system and then using it. Results show that there is a qualitative and quantitiative improvement of the relations extracted.

Keywords: Semantc Relation Extraction, Anaphora Resolution, Information Retrieval, Natural Language Processing

1 Introduction and Background

Semantic relations, as well as anaphoric expressions, are everywhere in natural language. The former are information that links entities and the latter are linguistic expressions referring to entities in text. Semantic Relation Extraction (SRE) is the task of retrieving semantic relations correctly and Anaphora Resolution (AR) is the task of linking expressions referring to entities in text. Recent works such as [25] and [3] showed that semantic information helps both tasks. In this paper we use AR to increase the information about entities in order to improve the performance of SRE in Italian and collect information for the Livememories Project[1], a database of the collective memory of the Italian region Trentino Alto Adige. While in English there is a lot of effort in the integration of SRE and AR (see for example [11] and [6]), there is very little in Italian.

We developed three systems based on three different extraction techniques: the first one is based on **Supervised relation extraction** (see [7]), that exploits manually labeled corpora with semantic relations decided a-priori and learning algorithms in order to classify relations using features such as entity types, work patterns, part-of-speech and the like. The second system is based on **Open relation extraction** (see [2]), that extracts tuples of entities and relational patterns

[1] http://www.livememories.org

by means of string matching and constraints. The third system, **Knowledge-Base relation extraction**, that is the first step of distant supervision (used for example by [12], [22] and [13] among others), exploits named entities in order to retrieve relations from knowledge bases. Each approach has its stength points and its weaknesses, for example the supervised approach yields usually high performances but requires the definition of a finite set of semantic relation classes, that is impossible to define once and for all, expecially for different languages. The open relation extraction approach is powerful under the aspect of the variety of semantic relations that can be extracted, but has the disadvantage that relation classes are expressed by string patterns, hence it is difficult to generalize from them. Knowledge-base relation extraction is bound to the existance and availability of knowledge bases in the desired language and requires word sense disambiaguation.

The Anaphora Resolution system instead groups entities in coreference chains, and this provides much information than having single entities. For example we can count the number of mentions in the chains and this gives us a measure of how much the entity is important in the text.

We tested the improvement in SRE running the three different systems and comparing the relations extracted by each system between entities and between coreference chains. Each system extracts different kind of relations in order to fill a filecard about each entity with as much information as possible.

In the next section we present the systems and the dataset in detail. In the third section we report the results and we draw some provisional conclusions.

2 Dataset and Relation Extraction Systems

2.1 Building the Dataset

We sampled a very small developement set (about 400 tokens), a training (about 20000 tokens) and a test set (about 5000 tokens) from L'Adige, a local news corpus used in the Livememories Project. We annotated the datasets with named entity types (Person, Location, Organization, Geopolitical Entity), Part-of-Speech tags, lemmas and End-Of-Sentence markers using TextPro (see [16]). Then we processed the dataset with Bart (see [17]), that provided automatic annotation of coreference chains, and with the italian Wikimachine[2] (see [8]) that provided links for the entities that have an entry in Wikipedia. We automatically annotated and manually corrected the test set with the gold standard relations for all systems. The observed agreement between the annotators and three Amazon's Mechanical Turk raters is k=0.9 (see [5] and [1]).

2.2 Building Relation Extraction Systems

Supervised Relation Extraction System For the supervised system we defined semantic relations as underspecified information between pairs of entities

[2] http://thewikimachine.fbk.eu/

in a sentence. We extracted all the entity type pairs from the training set (reported in table 1). Partly following the Relation Detection and Characterization task[3] (RDC, see [4]) we selected the following semantic relation classes, adapted to the entity types we had in our training set: We developed a classifier only for the relations whose frequency is above 10%, the classes are defined as follows:

Table 1. entity type pairs and frequency

Class	frequency
Person-Person (pp)	39,629%
Person-Organization (po)	21.291%
Person-GeoPolitical (pg)	15.172%
GeoPolitical-GeoPolitical (gg)	10.907%
Organization-GeoPolitical (og)	8.978%
GeoPolitical-location (gl)	1.857%
Person-location (pl)	1.238%
Organization-location (ol)	0.928%

– **Person-Social (SOCpp)**: one Person has a family relationship with another Person.
– **GeoPolitical-Affiliation (GAFpg)**: one Person lives, works in, or is citizen of a GeoPoliticalEntity.
– **Employment (EMPpo)**: one Person works for an Organization.
– **Physical-Location (PHYgg)**: one GeoPoliticalEntity is near or included in the other GeoPoliticalEntity.

We annotated those relations in the training set and we developed one binary classifier for each one using Support Vector Machines ([15]) in Weka ([24]). Given as features named entity types, punctuation, distance between the two entities and number of mentions (in the case we have coreference information) the classifier predicts whether or not there is the expected semantic relation class for each entity pair. For example for a pair of the type 'Person-Person" the classifier predicts if there is the 'SOCpp" relation or not. The performance of the supervised system is reported in table 2

Table 2. supervised classifiers' performance

Class	avg P	avg R	avg F1
SOCpp	0.891	0.875	0.826
GAFpg	0.738	0.74	0.721
EMPpo	0.615	0.617	0.616
PHYgg	0.657	0.58	0.557

[3] http://projects.ldc.upenn.edu/ace/docs/EnglishRDCV4-3-2.PDF

Open Relation Extraction System In the open extraction system semantic relations are defined as valid patterns found in the context between pairs of entities. Partly following Banko and Etzioni [2] we consider valid patterns the ones that include at least one verb or one preposition, that are not between self-relational pairs (pairs with two mentions of the same entity) and that follow some constraints. In order to select the constraints we run an experiment to check the parameters that affect the interpretation of patterns of semantic relations in Italian. In particular we tested long distance patterns (LD), found between all entities in a sentence; short distance patterns (SD), patterns found between nearest entities; and barriers (b), punctuation such as commas, parentheses and hyphens in the context between entities). We extracted 40 sentences from devset1: 10 short distance relations with barriers (bSD), 10 short distance without barriers (SD), 10 long distance with barriers (bLD), 10 long distance without barriers (LD). We run this experiment using Amazon mechanical Turk, which has been already exploited successfully for a number of tasks including Word Sense Disambiguation, Recognizing textual Entailment (see [20]) and Semantic Relation Extraction (see [9]). We asked raters to judge the relation patterns we extracted. For example in the sentence "*Ci sono 100000 DONNE che lavorano in TRENTINO* (There are 100000 WOMEN working in TRENTINO)" the extracted relation pattern is "*lavorare.in(donne,Trentino)* work.in(women,Trentino)". We gave four possible answers to the raters:

- Y) yes, the relation is correct;
- O) there is a relation but it is expressed by another pattern in the sentence;
- X) there is relation but it is not expressed in the sentence;
- N) there is no relation at all between the entities.

The agreement between the raters, is k=0.428 [5]. We kept only the examples where raters agreed. The results, reported in table 3, suggest that patterns between short distance pairs are more likely to be interpreted as relations and that barriers block the interpretation of semantic relations. We decided to extract patterns of verbs, nouns, prepositions and negative particles between nearest pairs of entities that are not blocked by barriers and that have at least one verb or one preposition.

Table 3. Agreed examples and parameters.

Parameters	**Y**	**O**	**X**	**N**
bSD	2	0	0	0
SD	3	0	0	0
bLD	0	0	0	2
LD	2	0	0	0

Knowledge-Base Relation Extraction System For this task we define semantic relations as the fields in the Wikipedia infoboxes. Related entities are the

title of wikipedia page and the value in a specific field of the infobox. The extraction system calls the Wikipedia Application Programming Interface[4] (APIs) by means of the link provided by the Wikimachine. Using Wikipedia APIs has the advantage to be always up-to-date, unlike dumps.
In order to decide which relations to extract from the infobox we tested two sets of infobox fields:

- set1
  headquarter(PHY),
  profession (ISA),
  political party+club+group (EMP)
- set2
  headquarter (PHY),
  profession (ISA),
  president(EMP),
  mayor (GAF)

The results of this experiment, reported in table 4, show clearly that set2 yield better performance than set1.

Table 4. Field-relation selection.

Fields	**P**	**R**	**F1**
set1	0.400	0.143	0.211
set2	1	0.263	0.416

Interaction between Anaphora Resolution and the Systems The three relation extraction systems follow the pipeline represented in figure 1.
In the Entity and mention extraction step we collect two variables: **m** and **r**. Variable **m** is the progressive mentions per entity count, a numeric variable that gives information about the importance of the entity in the text, and **r**, which is returned by a function that checks if the entity type is the same for all the mentions in a coreference chain, it is a binary variable. We use **r** as filter for the knowledge-base relation extraction system and **m** as filter the open relation extraction system. The supervised system exploits **m** as a feature in the Support Vector Machine model.

3 Results and Discussion

We run all the systems on the test set first using only entities (ent) and then using coreference chains (coref). As evaluation measures we provide the entity-relation ratio (named entity-relation ratio for the knowledge-base extraction system),

[4] `http://it.wikipedia.org/w/api.php`

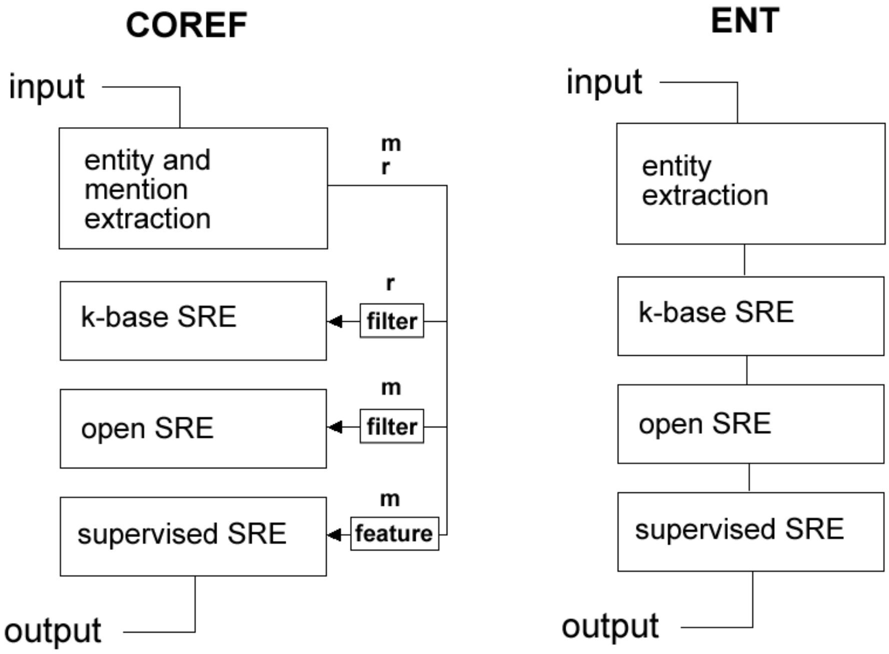

Fig. 1. Pipeline of SRE systems.

that tells us the amount of relations extracted, given the entities, and F1 score, that measures the relations retrieved correctly. Entity relation ratio is defined in the formula 1, where e is the entity count and r is the relation count

$$\frac{r}{e} \tag{1}$$

and F1 score is defined in 2, where p is precision and r is recall.

$$2\frac{pr}{p+r} \tag{2}$$

Results, reported in table 5, show an overall improvement in the entity-relation

Table 5. Systems Results

systems	ratio(ent)	ratio(coref)	F1(ent)	F1(coref)
knowledge-base	0.325	0.473	0.781	0.786
open	0.109	0.159	0.512	0.527
supervised	0.094	0.136	0.5	0.548

ratio due to the fact that coreference chains link entities and the relation between them. The improvement is larger for the knowledge-base extraction system (0.148), while for the open (0.05) and the supervised (0.042) systems the improvement is smaller. The opposite is true for the improvement in F-measure: the supervised system, that exploits the **m** feature in a support vector machine moodel is the one that yield the best improvement, while knowledge-based and open relation extraction, that use the **r** and **m** features respectively as filters, obtained smaller improvements.

4 Conclusions and Future Work

In this paper we presented three different systems for Semantic Relation Extraction that work jointly thanks to an Anaphora Resolution system. The three different systems are based on different extraction techniques and are trained for Italian language, for which the effort in the field of Semantic Relation Extraction is yet limited. To obtain this result we produced a dataset, which is small but very precious for the research community since it can be used as a benchmark for relation extraction in Italian. For the future we wish to work on the integration of different SRE techniques in an hybrid semantic relation extraction system and test iterative ways to improve anaphora resolution with semantic relations in Italian.

Acknowledgements This work has been realized also thanks to the support from the Provincia autonoma di Trento and the Fondazione Cassa di Risparmio di Trento e Rovereto. We wish to thank Christian Girardi, Claudio Giuliano and Olga Uryupina for making available to us the corpus L'Adige, the Wikimachine and BART.

References

1. Artstein R., Poesio M.. Intercoder agreement for Computational Linguistics. In Computational Linguistics, v. 34(4): 555–596. (2008).
2. Banko, M., Etzioni, O. The tradeoffs between traditional and open relation extraction. In Proceedings of ACL, (2008).
3. Bryl, V., Giuliano, C., Serafini, L., Tymoshenko, K.: Supporting Natural Language Processing with Background Knowledge: Coreference Resolution Case. In Proceedings of ISWC2010. (2010).
4. Doddington, G., A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel . The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. *Proceedings of LREC.* (2004)
5. Fleiss, J. L. Measuring nominal scale agreement among many raters. In Psychological Bulletin, Vol. 76, No. 5 pp. 378-382. (1971)
6. Gabbard, R., Freedman, M. Weischedel, R. Coreference for Learning to Extract Relations: Yes, Virginia, Coreference Matters. In Proceedings of ACL 2011, Portland. (2011)

7. Girju, R., Badulescu, A., Moldovan, D.: Automatic Discovery of Part-Whole Relations. In Computational Linguistics, 32(1), pp. 83–136. (2006).
8. Giuliano, C. Gliozzo, A. M., Strapparava, C. Kernel Methods for Minimally Supervised WSD. In Computational Linguistics, 35(4), pp. 513-528. (2009).
9. Gormley, M. R., Gerber, A., Harper, M., Dredze, M. Non-expert correction of automatically generated relation annotations. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. (2010).
10. Hearst, M. Automatic Acquisition of Hyponyms from Large Text Corpora: In Proceedings of COLING-92, (1992).
11. Ji, H., Westbrook, D. Grishman, R. Using Semantic Relations to Refine Coreference Decisions. In Proceedings of HLT/EMNLP 05, Vancouver. (2005)
12. Mintz, M., S. Bills, R. Snow, D. Jurafsky. Distant supervision for relation extraction without labelled data. In Proceedings of ACL-IJCNLP. (2009).
13. Nguyen, T. V., Moschitti, A. End-to-End Relation Extraction Using Distant Supervision from External Semantic Repositories. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. (2011)
14. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In Proceedings of Coling/ACL-06, (2006).
15. Platt, J. Machines using Sequential Minimal Optimization. In Schoelkopf, B., Burges, C., Smola, A. (ed), *Advances in Kernel Methods, Support Vector Learning.* (1998).
16. Pianta, E., Girardi, C., Zanoli. R.: The TextPro tool suite. In Proceedings of LREC, (2008).
17. Poesio, M., Uryupina, O., Versley, Y.: Creating a Coreference Resolution System for Italian. In Proceedings of the Seventh conference on International Language Resources and Evaluation, (2010).
18. Reed, S., Lenat, D. Mapping Ontologies into Cyc. In Proceedings of AAAI 2002 Conference Workshop on Ontologies For The Semantic Web. Edmonton. (2002)
19. Sanchez O., Poesio M., Kabadjov M.A., Tesar R.: What kind of problems do protein interactions raise for anaphora resolution? A preliminary analysis. In Proceedings of second symposium on Semantic Mining in Biomedicine, (2006),
20. Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y. Cheap and fast, but is it good?: evaluating non-expert annotations for natural language tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu. (2008)
21. Fabian M. Suchanek, F., M. Kasneci, G. Weikum, G. Yago: A Core of Semantic Knowledge. In Proceedings of 16th international World Wide Web conference, Banff. (2007)
22. Tymoshenko, K., Giuliano, C. FBK-IRST: Semantic Relation Extraction using Cyc,2010, In Proceedings of 5th International Workshop on Semantic Evaluations (SemEval-2010), (2010).
23. Turney, P.D.: Expressing implicit semantic relations without supervision, In Proceedings of Coling/ACL-06, (2006).
24. Witten, I, H., Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques with Java implementations. Morgan and Kaufman, San Francisco, CA. (2005).
25. Xu, F. Uszkoreit, H. Li, H.: Task Driven Coreference Resolution for Relation Extraction. In Proceeding of ECAI, (2008).