Entity Mention Detection using a Combination of Redundancy-Driven Classifiers

Silvana Marianela Bernaola Biggio, Manuela Speranza, Roberto Zanoli
Fondazione Bruno Kessler
Via Sommarive 18, Povo, 38123 Trento, Italy
E-mail: bernaola, manspera, zanoli{@fbk.eu}

Abstract

We present an experimental framework for Entity Mention Detection in which two different classifiers are combined to exploit Data Redundancy attained through the annotation of a large text corpus, as well as a number of Patterns extracted automatically from the same corpus. In order to recognize proper name, nominal, and pronominal mentions we not only exploit the information given by mentions recognized within the corpus being annotated, but also given by mentions occurring in an external and unannotated corpus. The system was first evaluated in the Evalita 2009 evaluation campaign obtaining good results. The current version is being used in a number of applications: on the one hand, it is being used in the LiveMemories project, which aims at scaling up content extraction techniques towards very large scale extraction from multimedia sources. On the other hand, it is being used to annotate corpora, such as Italian Wikipedia, thus providing easy access to syntactic and semantic annotation for both the Natural Language Processing and Information Retrieval communities. Moreover, a web service version of the system is available and the system is going to be integrated into the TextPro suite of NLP tools.

1. Introduction

We present an experimental framework for Entity Mention Detection, which exploits Data Redundancy attained through the annotation of a large text corpus, and a number of Patterns extracted automatically from the same corpus.

Entity Mention Detection (EMD) is an extension of Named Entity Recognition (NER), the task of recognizing different types of entities mentioned in a text. While NER systems are required to identify only proper name mentions, EMD systems are expected to deal with additional mention levels (categories). The system we present recognizes proper name mentions (NAM), e.g. “Alessandro Delpiero”, nominal mentions (NOM), e.g. “the player”, and pronominal mentions (PRO), e.g. “he”. It distinguishes four types of entities: Person (PER), e.g. “Alessandro Delpiero”, Organization (ORG), e.g. “Juventus”, Location (LOC), e.g. “Mount Everest”, and Geo-Political Entities (GPE), e.g. “Torino”.

The system draws from two different systems we developed earlier, one that took part in the Automatic Content Extraction 2008 evaluation for English (ACE08[1]) and one that was ranked highest in the Named Entity Recognition task at the Evalita 2009 evaluation campaign for Italian (Zanoli et al., 2009).

In those two systems, two new types of features were introduced to recognize NAM mentions:

- Data Redundancy, which is attained when the same entity occurs in different places in different documents.
- Patterns, i.e. 3-grams, 4-grams and 5-grams preceding and following recognized mentions.

The present work continues along the same lines and shows that Data Redundancy, and to a lesser extent Patterns, are useful in recognizing not only NAM but also NOM and PRO mentions.

The system has been evaluated in the Evalita 2009 evaluation campaign obtaining good results. Currently, it is being applied in LiveMemories, a project that aims at scaling up content extraction techniques towards very large scale extraction from multimedia sources, as well as for the annotation of Italian Wikipedia. Furthermore, a web service version of the system is now available and the system is going to be integrated into TextPro (Pianta et al. 2008), a collection of NLP tools whose functions range from tokenization to PoS tagging, NER, and cross-document co-reference.

The paper is structured as follows. Section 2 describes some related work; Section 3 describes the system’s architecture; Section 4 presents the results obtained by the system at Evalita 2009 and some experiments we made to evaluate the impact of Data Redundancy; Section 5 describes two current applications of the system; Section 6 describes the system as a web service application; finally, in Section 7, we draw some conclusions

2. Related work

Spurred on by the Message Understanding Conferences (MUC), several evaluation campaigns have been organized (among these, CoNLL[2] and ACE[3] are probably the most competitive and prestigious) and a considerable amount of work has been undertaken on NER in recent years. There are two main approaches to NER: one is based on Machine Learning methods (such as Support Vector Machines, Hidden Markov Models, and Maximum

[1] http://www.itl.nist.gov/iad/mig/tests/ace/2008/

[2] http://cnts.uia.ac.be/conll2002/ner/

[3] http://www.nist.gov/speech/tests/ace/

Entropy) and exploits a set of features (e.g. Part of Speech, orthographic features, information about the context in which a word appears, etc.) to perform statistical classifications; the other one is based on knowledge-based techniques and uses a set of hand-written rules in order to implement a specific grammar for named entities.

A number of systems have been developed for each approach: concerning the knowledge-based systems, it is worth mentioning the work done by Maynard et al. (2001) and Arevalo et al. (2004); as for machine learning systems, Carreras et al. (2002), Mayfield et al. (2003), and Florian et al. (2003).

Further work on NER tried to exploit global information (i.e. the fact that the same entity may occur more than once in the same document); for example, Mikheev et al. (1998) used information from the whole document, Borthwick (1999) proposed an additional system based on maximum entropy, trying to correct mistakes by using co-reference resolution (i.e. finding mentions referring to the same entity), and Chieu and Ng (2003) exploited information from the whole document to classify words using just one classifier.

3. System architecture

Our system is composed of the combination of two classifiers in cascade:

- A first classifier based on YamCha[4], an open source implementation of Support Vector Machines (SVMs);
- A second classifier based on disambig [5], an HMM-based tagger.

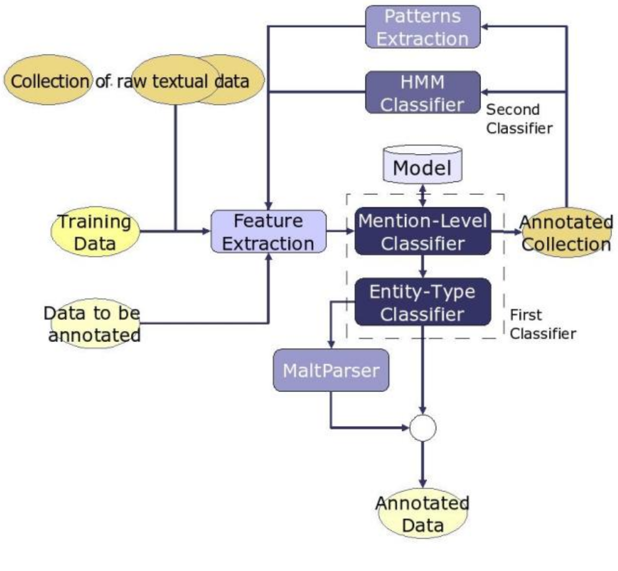

Figure 1: The system architecture

Figure 1 shows the system architecture; the system runs in two main phases.

In the first phase, the first classifier trained on manually annotated data (*training data*) is used to identify NAM, NOM and PRO mentions in an external *collection of raw textual data* (approximately one billion words) extracted from local and national newspapers as well as from Italian Wikipedia. The annotation produced by the first classifier (*annotated collection*) is used to extract a number of patterns and to train the second classifier. Then, the second classifier performs a new annotation of the *training data* and a preliminary annotation of the *data to be annotated*.

In the second phase, the first classifier performs the final annotation of the *data to be annotated* exploiting Patterns and Data Redundancy as two additional features and produces the *annotated data*.

It should be noted that the first classifier also consists of two classifiers in a cascade: (i) a mention level classifier, which identifies the syntactic head of a mention and its mention level and (ii) an entity type classifier which recognizes its type. In addition, in order to recognize the extension of a mention we used an implementation of MaltParser for Italian that took part at Evalita 2009 in the Dependency Parsing task (Lavelli et al. 2009).

3.1 Patterns

We considered as candidate patterns all 3-grams, 4-grams and 5-grams that precede and follow each recognized mention in the *annotated collection*.

Based on the intuition that certain patterns are more related to a type of mention, we used a formula based on TF-IDF (Term Frequency – Inverse Document Frequency) to select the patterns for each category (GPE, LOC, ORG, PER).

For each pattern *pk*, we calculated the Pattern Frequency for each category *cj*; see Formula 1, where #(*pk*, *cj)* denotes the number of times the pattern *pk* appears with mentions belonging to the category *cj*. Then, we calculated the *Pattern Frequency Inverse Category Frequency (pficf)*; see Formula 2, where #C(*pk*) denotes the number of categories with which the pattern *pk* appears at least once. This formula is based on:

(a) the *pattern frequency assumption*: the more frequently the pattern *pk* occurs with the category *cj*, the more important it is to *cj*.

(b) the *inverse category frequency assumption*: the more categories the pattern *pk* occurs with, the smaller its contribution is in characterizing the semantics of a category which it co-occurs with.

$$\mathrm{pf}\,(pk, cj) = \begin{cases} 1 + \log \#(pk, cj), & \text{if}\#(pk, cj) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Formula 1. Probability that a pattern co-occurs in a particular category

$$\mathrm{p}ficf(pk, cj) = \mathrm{pf}\,(pk, cj) * \log |C| /\#\mathrm{C}(pk)$$

Formula 2. TF-IDF weight of a pattern with a certain category

Finally, for each category, we ranked all candidate

[4] http://chasen.org/~taku/software/YamCha

[5] http://www.speech.sri.com/projects/srilm/

patterns on the basis of p*ficf* and arbitrarily selected the top 400,000 patterns. Table 1 shows some examples of the top 3-gram patterns for the GPE, ORG and PER mention types.

GPE	ORG	PER
il sindaco di <x>	il presidente del <x>	a cura di <x>
del comune di <x>	il leader della <x>	in memoria di <x>
la citta` di <x>	Da parte della <x>	il marito di <x>
nei pressi di <x>	il segretario del <x>	<x> ha detto che

Table 1: Top 3-gram patterns for GPE, ORG and PER

3.2 Data Redundancy

Sometimes a mention appears in an ambiguous context and it is difficult for the classifier to guess which is its correct type. In this case, it would be helpful to know the probability that a mention belongs to a certain category. E.g. if the mention "*ROMA*" appears multiple times in a corpus and the classifier tags it as GPE 70% of the time, as ORG 20% of the time, and 10% of the time it is tagged as not belonging to a category; then, it is more probable that when having a new document and the mention "*ROMA*" in it, it would be tagged as GPE. In this sense, we use a classifier to recognize all the mentions in a large corpus, in order to obtain the probability distribution for all mentions across all categories.

4. Evaluation and feature analysis

4.1 Evaluation of the EMD system's

The EMD system we present was evaluated in the EVALITA 2009 EMD task (Bernaola et al. 2009), a subtask of the Local Entity Detection and Recognition (LEDR) task. The organizers report a Value score (as defined for the ACE 2008 evaluation campaign) of 65.7% for Entity Mention Detection (Bartalesi e Sprugnoli, 2009).

The LEDR task at EVALITA 2009 consisted of (i) the detection of entity mentions in a corpus, assigning to each of them a mention level (i.e. NAM, NOM or PRO), type (PER, ORG, LOC or GPE) and corresponding subtype (see Table 2), and (ii) the recognition of all mentions referring to each entity (Local co-reference subtask).

Type	Subtypes
GPE (Geo-Political Entity)	Continent, Country-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province
LOC (Location)	Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body
ORG (Organization)	Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports
PER (Person)	Group, Indeterminate, Individual

Table 2: List of entity types and subtypes

4.2 Feature analysis

In the following paragraphs, we explain the experiments we conducted in order to analyze the influence of some of the features used in the first classifier.

Since our goal was to evaluate, most of all, the importance of Data Redundancy, we built a basic classifier in order to recognize the head of the mentions, using as features the token, Part of Speech (PoS), the prefix (first 3 characters of the token), suffix (last 3 characters of the token) and orthographic features (such as *is_upper_case*, *is_mixed_case*, *is_capitalized*, *is_abbreviation*, among others) in a window context of size 5. After that, we tagged a large external corpus in order to obtain the Redundancy and Pattern Features. Once these features were obtained, we carried out 14 experiments in order to evaluate the importance of each of the features shown in Table 3.

Features	Classifier													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Token	X	X	X	X	X	X	X	X	X	X	X	X	X	X
POS		X							X	X	X	X	X	X
Prefix-suffix			X					X		X	X	X	X	X
Orthographic				X				X	X		X	X	X	X
Redundancy					X			X	X	X		X	X	X
Pattern						X		X	X	X	X		X	X
Gazetteers							X	X	X	X	X	X		X

Table 3: List of features used in each classifier

These experiments are based on I-CAB, the Italian Corpus Annotation Bank, which consists of 525 news stories taken from four different days of the local newspaper "L'Adige" (Magnini et al. 2006); in particular we used three days as training data (September 7th, 2004; September 8th, 2004 and October 7th, 2004) and one day as test data (October 8th, 2004).

Table 4 presents the general and class-specific FB1 measure for each experiment. Since there are not many examples of pronouns (PRO) that refer to Geopolitical Entities (GPE), Locations (LOC) and Organizations (ORG); we will reduce the analysis to PROs that make reference to Persons (PER).

With respect to the Part of Speech feature (PoS); we can compare the results of experiment 14, which takes into account all the features, and experiment 8 that omits this feature. The general FB1 score does not change significantly and the same happens with each particular class. Something similar happens with the prefix-suffix and orthographic features (see experiments 9 and 10).

However, the results change when the Data Redundancy feature is not taken into account. According to experiment 11, the FB1 measure decreases almost 5% with respect to the 14th experiment. Looking in detail at each class, we can see that as expected, it is very helpful for NAM classes; for nominal names it seems that the most significant change is for NOM-GPE which suffers a decrease of around 20% when this feature is not used; for

	Only Token	Adding one feature to the token						Taking out 1 feature						All
Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14
General	59,81%	63,58%	59,51%	65,26%	78,15%	60,67%	69,82%	79,20%	79,97%	80,06%	74,09%	79,28%	78,70%	79,58%
NAM_GPE	66,96%	67,40%	65,94%	71,94%	80,51%	67,54%	75,72%	82,78%	84,30%	83,23%	78,37%	82,83%	81,99%	83,65%
NAM_LOC	63,55%	61,82%	67,26%	66,10%	71,19%	64,81%	69,35%	71,88%	75,38%	73,02%	77,52%	73,02%	69,92%	73,02%
NAM_ORG	46,11%	52,39%	43,00%	52,98%	70,12%	45,62%	60,72%	73,09%	74,77%	74,36%	66,81%	72,94%	72,65%	73,92%
NAM_PER	55,66%	72,54%	59,91%	72,96%	88,14%	59,64%	82,02%	91,31%	92,22%	91,91%	88,86%	92,03%	88,17%	91,63%
NOM_GPE	48,33%	47,06%	46,43%	49,61%	74,82%	45,90%	49,23%	75,00%	75,52%	73,76%	55,38%	75,18%	76,39%	75,86%
NOM_LOC	44,19%	34,67%	36,36%	57,14%	59,77%	39,02%	47,31%	59,79%	59,57%	61,70%	55,10%	59,18%	64,65%	62,37%
NOM_ORG	58,65%	51,33%	52,31%	58,40%	72,31%	57,83%	60,40%	71,70%	71,37%	71,89%	64,03%	70,41%	72,24%	71,46%
NOM_PER	70,18%	68,95%	70,36%	72,65%	85,63%	70,22%	75,48%	86,02%	86,15%	86,66%	78,29%	86,08%	86,61%	86,32%
PRO_GPE	0,00%	9,09%	0,00%	8,00%	9,52%	17,39%	15,38%	23,08%	23,08%	32,00%	14,29%	24,00%	32,00%	30,77%
PRO_LOC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	33,33%	0,00%	0,00%	0,00%
PRO_ORG	25,00%	29,53%	26,87%	29,73%	30,99%	25,95%	28,37%	31,72%	30,34%	30,77%	27,59%	30,56%	29,37%	29,17%
PRO_PER	66,33%	68,06%	65,71%	64,64%	72,46%	67,88%	65,57%	70,08%	70,46%	71,49%	68,43%	69,97%	70,09%	69,58%

Table 4: Experiment results

pronouns, an increase of 33% for PRO-LOC when not using the redundancy feature does not mean that the feature is not helpful for this class, but rather that the small number of instances of this class could not be recognized.

The second feature we introduce is Patterns, which, according to experiments 12 and 14, seems to make no difference in the general FB1 score. But for the NOM-LOC class, the FB1 score decreases around 3% when not taking this feature into account.

Finally, when removing the gazetteer feature (experiment 13), the general FB1 measure is practically unaltered. Moreover, it seems to have a small negative effect for NOMs. The list of gazetteers we used for PER, ORG, GPE and LOC entities were obtained automatically from several resources such as the Italian phone book and Wikipedia, among others. It seems that the benefit that Data Redundancy gives to the classifier can be replaced by this feature.

5. Applications

The current version of the system is being used in a number of applications. Among these it is worth mentioning the annotation of Italian Wikipedia as well as two corpora within the LiveMemories project.

5.1 LiveMemories

Nowadays people share and preserve their memories using applications such as Facebook, BBC Memoryshare, and Flickr, among others. In addition, there is a lot of information available on the web which could be exploited to enrich these memories. The LiveMemories project[6] aims to do this for two specific Italian corpora, (i) articles from the local newspaper *"L'Adige"* (620,641 articles from January 1st 1999 to October 15th 2009) and (ii) blogs posted by students living in the university residence *"San Bartolomeo"* (located in Trento, Italy).

	Corpus	
	"L'Adige"	**"San Bartolomeo"**
Tokens	257,255,240	2,974
Mentions	26,569,147	32,291
PER	16,736,169 (62.99%)	1,759 (58.36%)
ORG	5,692,208 (21.42%)	761 (25.25%)
GPE	3,475,461 (13.08%)	423 (14.03%)
LOC	665,309 (2.50%)	71 (2.36%)

Table 5: Data about the LiveMemories annotated corpora

The EMD system has been used to identify all the mentions occurring in these two corpora; the results of the process can be seen in Table 5. Approximately 26 million mentions where found in the "L'Adige" corpus, of which 63% belong to the PER type, 21% to ORG, 13% to GPE and less than 3% to LOC. A similar distribution of mention types follows the "San Bartolomeo" corpus: around 58% of the mentions detected belong to the PER type, 25% to ORG, 14% to GPE and almost 3% to LOC.

5.2 Italian Wikipedia

The importance of the electronic encyclopedia Wikipedia[7] is constantly increasing, not only as a source of information for human users, but also as a resource for NLP research. Atserias et al. (2008), for instance, annotated the Wikipedia in English and made it freely available; along the same lines, we have created SWiiT (Semantic WIkipedia for ITalian), which is distributed under a GNU license.

SWiiT is the Italian Wikipedia[8] annotated at five different levels: (i) basic NLP processing, (ii) entity mentions, (iii) entity subtypes, (iv) entity co-reference and (v) dependency parsing. The annotation of Wikipedia is still a work in progress; the first two levels of annotation have been completed and we are now working on the third level

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

[7] http://en.wikipedia.org

[8] http://it.wikipedia.org

(http://textpro.fbk.eu/resources/SWiiT.html).

Basic NLP processing. We downloaded the XML version of Italian Wikipedia (February 2010) and parsed it with Wikipedia Extractor[9]. The resulting pure text version consists of 594,336 articles, for a total of 9,985,745 sentences, which we processed using TextPro (Pianta et al. 2008). We performed tokenization, sentence splitting and PoS-tagging using the TANL tagset[10], designed according to the EAGLES guidelines and derived from the morpho-syntactic classification of the ISST corpus (Montemagni et al. 2003). In Table 6 we present the distribution of the Parts of Speech in SWiiT.

Nouns	57,504,251	(29.5%)
Verbs	21,287,049	(10.9%)
Adjectives	14,081,189	(7.2%)
Adverbs	6,663,142	(3.4%)
Conjunctions	6,915,957	(3.5%)
Prepositions	33,350,240	(17.1%)
Pronouns	5,422,390	(2.8%)
Articles	14,984,991	(7.7%)
Determiners	1,733,643	(0.9%)
Numerals	5,803,973	(3.0%)
Predeterminers	212,948	(0.1%)
Interjections	30,619	(0.0%)
Punctuation	27,034,643	(13.8%)
Other	218,198	(0.1%)
Total tokens	195,243,233	

Table 6: Part of speech distribution in SWiiT

Entity mention detection was carried out using the system described in this paper. Table 7 presents the distribution of entity mentions by type and by mention level.

NAM_GPE	1,664,265	(10,8%)
NAM_LOC	158,137	(1,0%)
NAM_ORG	1,743,849	(11,3%)
NAM_PER	3,980,580	(25,8%)
Total NAM	**7,546,831**	**(48.9%)**
NOM_GPE	588,661	(3.8%)
NOM_LOC	384,509	(2.5%)
NOM_ORG	650,478	(4.2%)
NOM_PER	3,505,696	(22.7%)
Total NOM	**5,129,344**	**(33.2%)**
PRO_GPE	25,400	(0.2%)
PRO_LOC	9,393	(0.1%)
PRO_ORG	116,551	(0.8%)
PRO_PER	2,615,910	(16.9%)
Total PRO	**2,767,254**	**(17.9%)**
Total mentions	**15,443,429**	

Table 7: Entity distribution in SWiiT

Entity subtype annotation is currently being performed using an SWM classifier (Bernaola et al. 2009).

Cross-document entity co-reference. In order to enrich SWiiT with information about co-reference between person entity mentions occurring in different documents, we will use the classifier developed by Popescu and Magnini (2007), which participated in the SEMEVAL 2007 Web People Search task obtaining the second best result among 16 systems with the following performance in terms of the harmonic mean of purity and inverse purity: $F\alpha=_{0.5}$ = 0.77 (purity=0,75 and inverse purity=0.80).

Dependency parsing. Syntactic analysis will be performed using MaltParser, a system for data-driven dependency parsing (Lavelli et al. 2009).

SWiiT data files have a tabular format with one token per line, where each line consists of five columns (the version currently distributed consists of the first three columns). The five columns contain respectively: (i) the token (a blank line marks end of sentences); (ii) the TANL Part of Speech tag; (iii) the entity tag (which includes entity type and mention level type) in the IOB format ("B" for words at the beginning of a mention, "I" for words inside a mention, and "O" for all other words); (iv) the entity subtype; and (v) the ID used to mark entity co-reference. Table 8 presents a sample annotation of SWiiT.

Le	RD	B-NOM-PER	Group	E1
Spie	S	I-NOM-PER		
tedesche	A	I-NOM-PER		
cercarono	V	O		
Di	E	O		
uccidere	V	O		
Stalin	SP	B-NAM-PER	individual	E2
,	FF	O		
Churchill	SP	B-NAM-PER	Individual	E3
E	CC	O		
Roosevelt	SP	B-NAM-PER	Individual	E4

Un	RI	O		
Lungo	A	O		
discorso	S	O		
Di	E	O		
Winston	SP	B-NAM-PER	individual	E3
Churchill	SP	I-NAM-PER		
[…]		O		
tenuto	V	O		
A	E	O		
Fulton	SP	B-NAM-GPE	pop.center	E5
(	FB	O		
Missouri	SP	B-NAM-GPE	Country	E6
)	FB	O		

Table 8: A sample from SWiiT.

6. Web service

We used Apache Axis to make our system available as a Web service. Axis[11] is an open source, XML based Web

[9] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

[10] http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset

[11] http://ws.apache.org/axis/ riferimento

service framework. It consists of a Java and C++ implementation of a SOAP server, and various utilities and APIs for generating and deploying Web service applications. Our Web service[12] allows users to submit a document (either in plain text format or as a tokenized text) and have it annotated with entity mentions following the IOB format.

7. Conclusion and future work

While we first introduced *Data Redundancy* and *Patterns* at Evalita 2009 for the NER task, in this paper we presented an experimental framework for Entity Mention Detection that is able to use the same types of features to annotate nominal and pronominal mentions as well.

During the development of the EMD system, we discovered that the classifier finds greater difficulties in recognizing pronominal mentions; for instance, the word *che* is a pronoun when it makes reference to a previous mentioned entity, and it should be annotated as a pronoun only if it makes reference to an entity that belongs to the list of entity types to annotate (PER, ORG, LOC, GPE). But the main issue with pronouns is related to the context; continuing with the previous example, if the word *che* refers to an entity that has been mentioned in a previous sentence, which is out of context for the classifier, it is very probable that the classifier does not recognize it as a pronoun.

On the other hand, there are a lot of mentions that are ambiguous and sometimes they can be disambiguated because of the context; for example, the word *Benetton* can be a person or an organization; however in the sentence: "*Benetton ha detto che* …" (*"Benetton has said that* …"), it is most probably that the mention *Benetton* refers to a Person, since it is an Italian surname and usually the pattern "ha detto che" usually appears with mentions of type PER. This is the reason why we introduced the pattern feature. However, the results obtained were not what we expected; i.e. it did not make any difference in the FB1 measure; we believe this is because of the threshold used to select the patterns for each category. In future work, we would like to find out how to select it in order to get the right patterns for each class.

Finally, according to the results obtained in the experiments, the Data Redundancy feature seems to be the most important feature; it improves FB1 around 5% which is not the case for any other feature, including the gazetteer feature. One surprising thing is the huge increase (around 20%) it provides to nominal names that refer to geopolitical entities.

8. Acknowledgements

This work has been partially supported by the project LiveMemories, funded by the Provincia Autonoma of Trento.

9. References

Arevalo, M., Civit, M., Marti, M.A. MICE: A module for Named Entity Recognition and Classification. *International Journal of Corpus Linguistics*, 9(1):53 – 68, March 2004.

Atserias, J., Zaragoza, H., Ciaramita, C., Attardi, G. (2008). Semantically Annotated Shanphot of the English Wikipedia. In *Proceedings of LREC 2008*, 28-30 May 2008, Marrakech, Morocco.

Bartalesi Lenzi, V., Sprugnoli, R. (2009). EVALITA 2009: Description and Results of the Local Entity Detection and Recognition (LEDR) task. In *Proceedings of Evalita 2009*, workshop held at AI*IA, 12 December 2009, Reggio Emilia, Italy.

Bernaola Biggio, S.M., Zanoli, R., Giuliano, C., Uryupina, O., Versley, Y., Poesio, M. (2009). Local Entity Detection and Recognition Task. In *Proceedings of Evalita 2009*, workshop to held at AI*IA, 12 December 2009, Reggio Emilia, Italy.

Borthwick, A. (1999). A Maximum Entropy Approach to Named Entity Recognition. Ph.D. thesis. Computer Science Department, New York University.

Carreras, X., Marques, L., Padro, L. (2002). Named Entity Extraction using Adaboost. In *Proceedings of CoNLL-2002*, pages 167–170, Taipei, Taiwan, 2002.

Chieu, N.L., Ng, H.T. (2003). Named Entity Recognition with a Maximum Entropy Approach. In *Proceedings of the seventh conference on Natural Language Learning at HLT-NAACL 2003* - Volume 4, Edmonton, Canada, 160–163.

Florian, R., Ittycheriah, A., Jing, H., Zhang, T. (2003). Named Entity Recognition through Classifier Combination. In Walter Daelemans and Miles Osborne, (Eds.), *Proceedings of CoNLL-2003*, Edmonton, Canada, 168-171.

Lavelli, A., Hall, J., Nilsson, J., Nivre, J. (2009). MaltParser at the EVALITA 2009 Dependency Parsing Task. In *Proceedings of Evalita 2009*, workshop held at AI*IA, 12 December 2009, Reggio Emilia, Italy.

Magnini, B., Cappelli, A., Pianta, E., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R., Romano, L., Girardi, C., Negri, M. (2006). Annotazione di contenuti concettuali in un corpus italiano: I-CAB. In *Proceedings of SILFI 2006*. Florence, Italy.

Mayfield, J., McNamee, P., Piatko, C. (2003). Named Entity Recognition Using Hundreds of Thousands of Features. In Walter Daelemans and Miles Osborne, (Eds.), *Proceedings of CoNLL-2003*, Edmonton, Canada, 184-187.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., Wilks, Y. (2001). Named Entity Recognition from Diverse Text Types. In R. Mitkov, N. Nicolov, G. Angelova, K. Bontcheva and N. Nikolov (Eds.), *Proceedings of the First Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, 2001.

Mikheev, A., Grover, C., Moens, M. (1998). Description of the LTG System Used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference*,

[12] http://textpro.fbk.eu/typhoon.html

Fairfax, Virginia, USA.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F.M., Mana, N., Pianesi, F., Delmonte, R. (2003). Building the Italian Syntactic-Semantic Treebank. In Abeillé (Ed.), *Building and Using Parsed Corpora*. Language and Speech series, Kluwer, Dordrecht, 189-210.

Pianta, E., Girardi, C., Zanoli, R. (2008). The TextPro tool suite. In *Proceedings of LREC 2008*, 28-30 May 2008, Marrakech, Morocco.

Popescu, O., and Magnini, B. (2007). Web People Search Using Name Entities. In *Proceedings of SemEval-2007 Workshop*, co-located with ACL 2007, Prague, CZ, 23-24 June 2007.

Speranza, M. (2009). The Named Entity Recognition Task at EVALITA 2009. In *Proceedings of Evalita 2009*, workshop held at AI*IA, 12 December 2009, Reggio Emilia, Italy.

Zanoli, R., Pianta, E., Giuliano, C. (2009). Named Entity Recognition through Redundancy Driven Classifiers. In *Proceedings of Evalita 2009*, workshop held at AI*IA, 12 December 2009, Reggio Emilia, Italy.