From Mentions to Ontology: A Pilot Study

Octavian Popescu, Bernardo Magnini, Emanuele Pianta, Luciano Serafini, Manuela Speranza and Andrei Tamilin, ITC-irst, 38050, Povo (TN), Italy

Abstract— **In this paper we propose a pilot study aimed at an in-depth comprehension of the phenomena underlying Ontology Population from text. The study has been carried out on a collection of Italian news articles, which have been manually annotated at several semantic levels. More specifically, we have annotated all the textual expressions (i.e. mentions) referring to Persons; each mention has been in turn decomposed into a number of attribute/value pairs; co-reference relations among mentions have been established, resulting in the identification of entities, which, finally, have been used to populate an ontology. There are two significant results of such a study. First, a number of factors have been empirically identified which determine the difficulty of Ontology Population from Text and which can now be taken into account while designing automatic systems. Second, the resulting dataset is a valuable resource for training and testing single components of Ontology Population systems.**

I. Introduction

In this paper we propose an empirical investigation into the relations between language and knowledge, aiming at the definition of a computational framework for automatic Ontology Population (OP) from text.

While Ontology Population from text has received an increasing attention in recent years (see for instance, Buitelaar et al. 2005), mostly due to its strong relationship with the Semantic Web perspective, very little has been done in order to provide a clear definition of the task and to establish shared evaluation procedures and benchmarks. In this paper we propose a pilot study aimed at an in-depth comprehension of the phenomena underlying Ontology Population from Text (OPTM). Specifically, we are interested in highlighting the following aspects of the task:

- What are the major sources of difficulty of the task?
- How does OP from text relate to well known tasks in Natural Language Processing, such as Named Entity Recognition?
- What kinds of reasoning capabilities are crucial for the task?
- Is there any way to simplify the task so that it can be addressed in a modular way?
- Can we devise useful metrics to evaluate system performance?

We addressed the above questions through a pilot study on a limited amount of textual data. We added two restrictions with respect to the general OP task: first, we considered textual mentions instead of full text; second, we focused on information related to PERSON entities instead of considering all possible entities (e.g. ORGANIZATION, LOCATION, etc.).

Mentions, as defined within the ACE (Automatic Content Extraction)[1] Entity Detection Task (Linguistic Data Consortium, 2004) are portions of text that refer to entities. As an example, given a particular textual context, the two mentions "*George W. Bush*" and "*the U.S President*" refer to the same entity, i.e. a particular instance of PERSON whose first name is "*George*", whose middle initial is "*W.*", whose family name is "*Bush*" and whose role is "*President of the U.S.*".

As for PERSON entities, they were selected for our pilot study because they occur very frequently in the news document collection we analyzed. Most of the results we obtained, however, are likely to be generalized over the other types of entities.

Given the above-mentioned restrictions, the contribution of this paper is a thorough study of Ontology Population from Textual Mentions (OPTM). We have manually extracted a number of relevant details concerning entities of type PERSON from the document collection and then used them to populate a small pre-existing ontology. This led to two significant results of such a study. First, a number of factors have been empirically identified which determine the difficulty of Ontology Population from Text and which can now be taken into account while designing automatic systems. Second, the resulting dataset is a valuable resource for training and testing single components of Ontology Population.

We show that the difficulty of the OPTM task is directly correlated to two factors: (A) the difficulty of identifying attribute/value pairs inside a given mention and (B) the difficulty of establishing co-reference between entities based on the values of their attributes.

There are several advantages of OPTM that makes it appealing for OLP. First, mentions provide an obvious simplification with respect to the more general task of Ontology Population from text (cfr. Buitelaar et al. 2005); in addition, mentions are well defined and there are systems for automatic mention recognition which can provide the input for that task. Second, since mentions have been introduced as an evolution of the traditional Named Entity Recognition task (see Tanev and Magnini, 2006), they guarantee a reasonable level of complexity, which makes OPTM challenging both for the Computational Linguistics and the Knowledge Representation communities. Third, there already exist data annotated with mentions, delivered under the ACE initiative (Ferro et al. 2005, Linguistic Data Consortium 2004), which make it possible to exploit machine learning approaches. The

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

availability of annotated data allows for a better estimation of the performance of OPTM; in particular, it is possible to evaluate the recall of the task, i.e. the proportion of information correctly assigned to an entity out of the total amount of information provided by a certain mention.

The paper is structured as follows. Section II provides some background on Ontology Population and reports on relevant related work; Section III describes the dataset of the PERSON pilot study and compares it to the ACE dataset. Section IV introduces a new methodology for the semantic annotation of attribute/value pairs within textual mentions. In section V we describe the Ontology we plan on using. Finally, Section VI reports on a quantitative and qualitative analysis of the data, which help determining the main sources of difficulty of the task. Conclusions are drawn in Section VII.

.

II. RELATED WORK

Automatic Ontology Population (OP) from texts has recently emerged as a new field of application for knowledge acquisition techniques (Buitelaar et al., 2005). Although there is no widely accepted definition for the OP task, a useful approximation has been suggested by (Bontcheva and Cunningham, 2003) as Ontology Driven Information Extraction with the goal of extracting and classifying instances of concepts and relations defined in an ontology, in place of filling a template. A similar task has been approached in a variety of similar perspectives, including term clustering (Lin, 1998; Almuhareb and Poesio, 2004) and term categorization (Avancini et al., 2003). A rather different task is Ontology Learning, where new concepts and relations are supposed to be acquired with the consequence of changing the definition of the Ontology itself (Velardi et al. 2005).

The interest in OP is also reflected in the large number of research projects which consider knowledge extraction from text a key technology for feeding Semantic Web applications. Among such projects, it is worth mentioning Vikef (Making the Semantic Web Fly), whose main aim is to bridge the gap between implicit information expressed in scientific documents and its explicit representation found in knowledge bases; and Parmenides, which is attempting to develop technologies for the semi-automatic building and maintenance of domain-specific ontologies.

The work presented in this paper has been inspired by the ACE Entity Detection task, which requires that the entities mentioned in a text (e.g. PERSON, ORGANIZATION, LOCATION and GEO-POLITICAL ENTITY) be detected. As the same entity may be mentioned more than once in the same text, ACE defines two inter-connected levels of annotation: the level of the entity, which provides a representation of an object in the world, and the level of the entity mention, which provides information about the textual references to that object. The information contained in the textual references to entities may be translated into a knowledge base, and eventually into an Ontology.

III. DATA SET

The input of OPTM consists of textual mentions derived from the Italian Content Annotation Bank (I-CAB), which consists of 525 news documents taken from the local newspaper 'L'Adige'[2], for a total of around 180,000 words (Magnini et al., 2006). The annotation of I-CAB has been carried out manually within the Ontotext project[3], following the ACE annotation guidelines for the Entity Detection task. I-CAB is annotated with expressions of type TEMPORAL_EXPRESSION and four types of entities: PERSON, ORGANIZATION, GEO-POLITICAL ENTITY and LOCATION. Due to the morpho-syntactic differences between the two languages, the ACE annotation guidelines for English had to be adapted to Italian; for instance, two specific new tags, PROCLIT and ENCLIT, have been created to annotate clitics attached to the beginning or the end of certain words (e.g. *<veder*[*lo*]*>/to see him*).

According to the ACE definition, entity mentions are portions of text referring to entities; the extent of this portion of text consists of an entire nominal phrase, thus including modifiers, prepositional phrases and dependent clauses (e.g.*<il* [*ricercatore*] *che lavora presso l'ITC-irst>/the resercher who works at ITC- irst*).

Mentions are classified according to four syntactic categories: NAM (proper names), NOM (nominal constructions), PRO (pronouns) and PRE (modifiers).

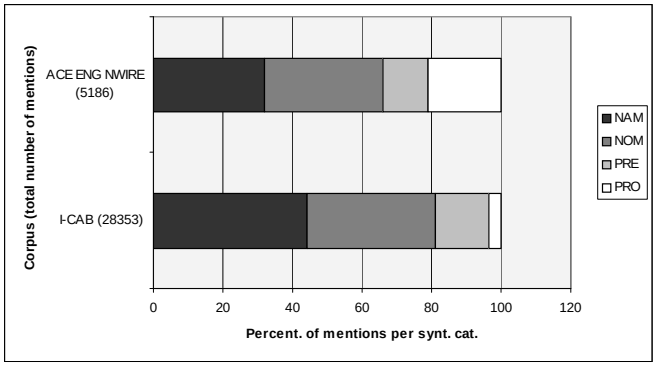

Fig. 1. Distribution of the four different ACE mention types in I-CAB and in the ACE 2004 Evaluation corpus (Newswire)

In spite of the adaptations to Italian, it is interesting to notice that a comparison between I-CAB and the newswire portion of the ACE 2004 Evaluation corpus (see Figure 1) shows a similar proportion of NAM and NOM mentions in the two corpora. On the other hand, there is a low percentage of PRO mentions in Italian, which can be explained by the fact that, unlike in English, subject pronouns in Italian can be omitted. As for the large difference in the total number of mentions annotated in the two corpora (22,500 and 5,186 in I-CAB and ACE NWIRE respectively), this is proportional to their size (around 180,000 words for I-CAB and 25,900 words for ACE NWIRE), considering that some of the ACE entities

[2] http://www.ladige.it/
[3] http://tcc.itc.it/projects/ontotext/index.html

(i.e. FACILITY, VEHICLE, AND WEAPON) are not annotated in I-CAB.

As shown in Figure 2, the two corpora also present a similar distribution as far as the number of mentions per entity is concerned. In fact, in both cases more than 60% of the entities are mentioned only once, while around 15% are mentioned twice. Between 10% and 15% are mentioned three or four times, while around 6% are mentioned between five and eight times. The fact that the percentage of entities mentioned more than eight times in a document is higher in the ACE corpus than in I-CAB can be partly explained by the fact that the news stories in ACE are on average slightly longer than those in ACE (around 470 versus 350 words per document).

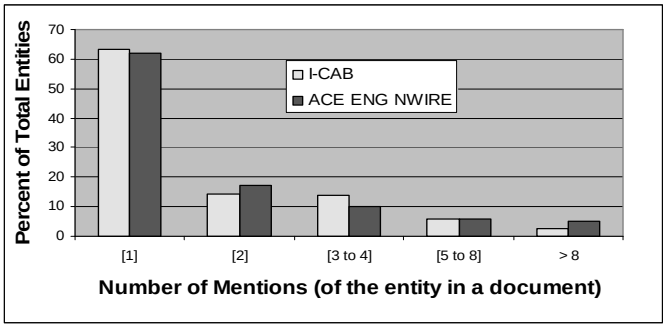

Fig. 2. Intra-document co-reference in I-CAB and in the ACE 2004 Evaluation corpus (Newswire)

IV. ATTRIBUTES for TYPE PERSON

After the annotation of mentions of type PERSON reported in the previous section, each mention was additionally annotated in order to individuate the semantic information expressed by the mention regarding a specific entity. As an example, given the mention "*the Italian President Ciampi*", the following attribute/value pairs were annotated: [PROVENANCE: *Italian*], [ROLE: *President*] and [LAST_NAME: *Ciampi*].

The definition of the set of attributes for PERSON followed an iterative process where we considered increasing amounts of mentions from which we derived relevant attributes. The final set of attributes is listed in the first column of Table 1, with respective examples reported in the second column.

A strict methodology is required in order to ensure accurate annotation. As general guidelines for annotation, articles and prepositions are not admitted at the beginning of the textual extent of a value, an exception being made in the case of the articles in nicknames (see Magnini et al., 2006B for a full description of the criteria used to decide on border cases).

Attributes can be grouped into bigger units, as in the case of the attribute JOB, which is composed of three attributes, ACTIVITY, ROLE, and AFFILIATION, which are not independent of each other. ACTIVITY refers to the actual activity performed by the person, while ROLE refers to the position they occupy. So, for instance, "*politician*" is a possible value of the attribute ACTIVITY, while "*leader of the Labour Party*" refers to the ROLE a person plays inside an organization. Each group of these three attributes is associated with a mention and all the information within a group has to be derived from the same mention. If different pieces of information derive from distinct mentions, we will have two separate groups. For instance, the three co-referring mentions "*the journalist of Radio Liberty*", "*the redactor of breaking news*", and "*a spare time astronomer*" lead to three different groups of ACTIVITY, ROLE and AFFILIATION. The obvious inference that the first two mentions belong conceptually to the same group is not drawn. This step is to be taken at a further stage.

attributes	**values**
FIRST_NAME	*Ralph, Greg*
MIDDLE_NAME	*J., W.*
LAST_NAME	*McCarthy, Newton*
NICKNAME	*Spider, Enigmista*
TITLE	*Prof., Mr.*
SEX	*actress*
ACTIVITY	*author, doctor*
AFFILIATION	*The New York Times*
ROLE	*manager, president*
PROVENANCE	*South American*
FAMILY_RELATION	*father, cousin*
AGE_CATEGORY	*boy, girl*
HONORARY	*the world champion 2000*
MISCELLANEA	*The men with red shoes*

Table 1. The attribute structure of PERSON

We started with the set of 525 documents belonging to the I-CAB corpus (see section III), for which we have manually annotated all PERSON entities (10039 mentions, see Table 2). The annotation individuates both the entities mentioned within a single document, called *document entities*, and the entities mentioned across the whole set of news stories, called *collection entities*. In addition, for the purposes of this work, we decided to filter out the following mentions: (i) mentions consisting only of one non-gender discriminative pronoun; (ii) nested mentions, i.e. in case inside a mention there is a smaller one, for example as in "*the president Ciampi"*, with "*Ciampi*" being the included one, only the largest mention was considered. In this way we obtained a set of 7233 mentions which represents the object of our study.

Number of documents	525
Number of mentions	10039
Number meaningful mentions	7233
Number of distinct meaningful mentions	4851
Number of document entities	3284
Number of collection entities	2574

Table 2. The PERSON Dataset

The average number of meaningful mentions for an entity in a certain document is 2.20, while the average number of distinct meaningful mentions is 1.47. However, the variation from the average is high, only 14% of document entities are

mentioned exactly twice. In fact, there are relatively few entities whose mentions in news have a broad coverage in terms of attributes, and there are quite a few whose mentions contain just the name. A detailed analysis is carried out in Section VI.

V. ONTOLOGY

The ontology adopted for the OPTM task is composed of two main parts. The first part mirrors the mention attribute structure and contains axioms (restrictions) on the attribute values. In this part, which we refer as the Entity T-Box (ET-box), we define three main classes corresponding to the three main entities, PERSON, ORGANIZATION and GEO-POLITICAL ENTITY. Each of these classes is associated with the mention attributes. An example of how the attributes are encoded in axioms in the ET-box is provided in Table 3.

ONTOLOGY AXIOM	**Encoded restriction**
PERSON ⊆(>0)HAS_FIRST_NAME	Every person has at least a first name
PERSON ⊆ (=1)HAS_LAST_NAME	Every person has exactly one last name
DOMAIN(HAS_FIRST_NAME) = PERSON	the first argument of the relation has_first_name must be a person
RANGE(HAS_PROVENANCE) = GEOPOLITICALENTITY	The second argument of the relation HAS_PROVENENCE must be a geopolitical entity

Table 3. Description of Ontology axioms

The second component of the ontology, called world knowledge (WK), encodes the basic knowledge about the world already available (see Table 4 for examples of axioms). This ontology has been semi-automatically constructed starting from the large amount of basic information available on the web. Examples of such knowledge are the sets of countries, main cities, country capitals, Italian municipalities, etc.

ONTOLOGY AXIOM	**Encoded restriction**
COUNTRY(Italy)	Italy is a country
HAS_CAPITAL(Italy,Rome)	Rome is the capital of Italy
CONTINENT ⊆ GEOPOLITICALENTITY	A country is a geopolitical entity
TOWN ⊆ GEOPOLITICALENTITY	A town is a geopolitical entity

Table 4. Description of Ontology axioms related to WK

As can be seen from the above examples, WK is composed of two types of knowledge: factual knowledge (the first two axioms in Table 4) and generic commonsense knowledge. The first type of knowledge can be obtained from the many ontological resources available on the web (see for instance swoogle.umbc.edu), while we have manually encoded the second in the ontology.

The process of OPTM combines the ontology ET-box with WK axioms and values of attributes recognized in textual mentions, and performs two main steps:

1. For each entry recognized in the text we create a new individual in the ontology, along with the individuals corresponding to the attribute values
2. We normalize the values by comparing the "string" values with the individuals present in the WK.

As an example of this process, consider the entry in Table 5.

FIRST_NAME	*Bob, B.*
LAST_NAME	*Marley*
PROVENANCE	*Caribbean*
ACTIVITY	*musician, guitar player*

Table 5. Attributes/Values examples

In the first phase we add the axioms in Table 6 to the ontology.

Person(person23)
HAS_FIRST_NAME(person23,first_name76)
HAS_LAST_NAME(person23,last_name93)
HAS_PROVENANCE(person23,geo_pol_entity35)
HAS_ACTIVITY(person23,activity43)
HAS_ACTIVITY(person23,activity44)
HAS_VALUE(first_name56, "Bob")
HAS_VALUE(first_name76, "B.")
HAS_VALUE(geo_pol_entity35, "Caribbean")
HAS_VALUE(activity43, "musician")
HAS_VALUE(activity44,"guitar player")

Table 6. Adding axioms to the Ontology

In the second phase, we attempt to match the values to the individuals in the WK and the Ontology is modified according to the result of the matching process. This process is based on the semantic matching approach described in (Bouquet, 2003).

In this phase the WK-part of the ontology take a crucial role. The main goal of this phase is to find the best match between the values of an attribute and the individuals which are already present in the WK A-box. This process can have two outputs. When a good-enough match is found between an attribute value and an individual of the WK A-box, then an equality assertion is added. Suppose for instance that the WK A-box contains the statement

STATE(Caribbean)

then the mapping process will find a high match between the value "Caribbean" (as a string) and the individual Caribbean (due to the syntactic similarity between the two strings, and the fact that both are associated to individuals of type GEOPOLITICALENTITY). As a consequence the assertion

Geo_pol_entity35 = Caribbean

is asserted in the A-box. Notice that the above assertion connects an individual of the WK with the value of an entity contained in the entity repository of the mentions.

When the mapping process does not produce a "good" mapping (where good is defined w.r.t., a suitable distance measure not described here) the value is transformed into an individual and added to the WK A-box. For instance, suppose that the mapping of the value "guitar player" will not produce a good matching value, then the new assertion

```
ACTIVITY(GuitarPlayer)
```

is added to the WK A-box and the assertion

```
activity44 = GuitarPlayer
```

is added to the A-box that links WK with the A-box of the mentions.

pairs inside a given mention and (B) the difficulty of establishing the co-reference of entities based on the values of their attributes.

In table 7 we find the distribution of the values of the attributes defined for PERSON. The first column lists the set of attributes; the second column lists the number of occurrences of each attribute, the third lists the number of different values that the attribute actually takes; the fourth column lists the number of collection entities which have that attribute. Using this table as base table we try to determine the parameters which give us no clues on the two factors above

.

Attribute	Occurrence of attribute in mentions	Different values for attribute	Collection entities with attribute	Distinct values within distinct mentions	Variability of values in attribute
FIRST_NAME	2299 (31%)	676	1592	13%	29%
MIDDLE_NAME	110 (1%)	67	74	1%	60%
LAST_NAME	4173 (57%)	1906	2191	39%	45%
NICKNAME	73 (1%)	44	41	0%	60%
TITLE	73 (1%)	25	47	0%	34%
SEX	3658 (50%)	1864	1743	38%	50%
ACTIVITY	973 (13%)	322	569	6%	33%
AFFILIATION	566 (7%)	389	409	8%	68%
ROLE	531 (7%)	211	317	4%	39%
PROVENANCE	469 (6%)	226	367	4%	48%
FAMILY_RELATION	133 (1%)	46	94	0%	34%
AGE_CATEGORY	307 (4%)	106	163	2%	34%
HONORARY	69 (0%)	63	53	1%	91%
MISCELLANEA	278 (3%)	270	227	5%	97%

Table 7. Distribution of values of attributes for PERSON

A. Difficulty of identifying attribute/value pairs

The identification of attribute/value pairs requires the correct decomposition of the mentions into non overlapping parts, each one carrying the value of one attribute. We are interested in estimating the distribution of attributes inside the mentions. Table 8 shows on the second and fourth columns the number of mentions which contain respectively 1, 2, 3, …, 12 attributes. As we can see, the number of mentions having more than 6 attributes is insignificant. On the other hand, the number of mentions containing more than one attribute is 3564, which represents 49,27% of the total, therefore one in two mentions is a complex mention. Usually, a complex mention contains a SEX value, therefore a two attribute mention practically has just one that might help in establishing co-reference. However, 92% of the mentions with up to 5 attributes, which covers 96% of all mentions, contain a NAME attribute, which, presumably, is an important piece of evidence in deciding on co-reference.

The difficulty of correct identification of the attribute/value pairs is directly linked to the complexity of a mention. Two values inside the mention belong to the same entity. Without recognizing the correct frontiers of a complex mention virtually 50% of the cases are treated badly.

#attributes	#mentions	#attributes	#mentions
1	3669 (50%)	7	34 (0,04%)
2	1292 (17%)	8	19
3	1269 (17%)	9	4
4	486 (6%)	10	4
5	310 (4%)	11	0
6	146 (2%)	12	0

Table 8. Number of attributes carried by mentions

attribute	2 attribute mention	3 attribute mention	4 attribute mention
FIRST_NAME	398	915	413
MIDDLE_NAME	5	20	34
LAST_NAME	467	1025	426
NICKNAME	27	16	2
TITLE	14	16	13
SEX	806	1240	501
ACTIVITY	273	135	413
AFFILIATION	82	91	80
ROLE	126	81	94
PROVENANCE	81	134	156
FAMILY_RELATION	76	24	103
AGE_CATEGORY	139	62	12
HONORARY	20	7	31
MISCELLANEA	80	59	11

Table 9. Distribution of attributes into complex mentions

A second difficulty of correctly identifying the attribute/value pairs comes from the combinatorial capacities of attributes inside a complex mention. If the diversity of attribute patterns in a complex mention is high, then the difficulty of their recognition is also high. Table 9 shows that the whole set of attributes is very well represented in the complex mentions and, interestingly, the number of attributes varies independently of the number of mentions, therefore their combinatorial capacity is high. The difficulty of their recognition varies accordingly.

The distribution of attributes inside mentions is presented in the second column of Table 7 in parenthesis. The figures give the probability that a person is mentioned by making reference to a certain attribute. For example, one may expect the LAST_NAME attribute to be present in 57% of mentions, and the NICKNAME attribute to be present in 0,001% of the total. In the fifth column we compute the same figures without repetition, considering the distinct values and distinct mentions. Considering also the figures that show the linguistic variability of values, we may obtain the probability of seeing a previously unseen value of a given attribute. The last column of Table 7 shows the variability of values for each attribute. For example, taking randomly a mention of FIRST_NAME, only in 29% of the cases that value is seen in the dataset just once.

The fifth column, distinct values within distinct mentions, and the sixth, variability of values in attribute, offer us insight into the difficulty of recognizing attribute/value pairs. The variability might be considered as representative of the amount of training a system needs in order to have a satisfactory coverage of cases. Intuitively, some of the attributes are close classes, while some other attributes, e.g. those who have name values, are open classes.

Probably, the importance of recognizing certain types of attributes is bigger than for other ones. If the occurrence of a new value of an important attribute is a rare event, a system must be very precise in catching these cases. We may assume that a high precision is more difficult to achieve than a lower one. The "distinct" column gives us a clue on this issue. For example, the relatively low figures for ACTIVITY, AFFILIATION, ROLE but their importance with respect to the OPTM task, tell us that sparseness could be an issue and therefore a precise system of their treatment must be used. Otherwise it will be hard to achieve the expected results.

Finally, we may notice that 39% of the mentions carry some other information than SEX and name related values, MISCELLANEA excluded. Therefore in all those cases the ontology is enriched with substantial information about the respective persons.

B. Difficulty of establishing Co-references among entities

The task of correctly identifying a value of a certain attribute inside a given mention is worth to be undertaken if the respective values play a role in other tasks, especially in the co-reference task. A relevant factor for co-reference is the perplexity of an attribute, i.e. the percentage of the entities characterized by a particular value, computed as the ratio between distinct values for a certain attribute and collection entities having that attribute (column III / IV in table 7). For example the perplexity of LAST_NAME is 14% (see Table 10). Therefore if we take randomly some values of LAST_NAME, 86% of them are pointing to just one person. In the case of SEX and MISCELLANEA, the perplexity is not defined.

attribute	perplexity
FIRST_NAME	58%
MIDDLE_NAME	10%
LAST_NAME	14%
NICKNAME	0%
TITLE	47%
SEX	-
ACTIVITY	44%
AFFILIATION	5%
ROLE	34%
PROVENANCE	52%
FAMILY_RELATION	39%
AGE_CATEGORY	35%
HONORARY	0%
MISCELLANEA	-

Table 10. Perplexity of PERSON attributes

By comparing the perplexity of LAST_NAME and MIDDLE_NAME one might erroneously conclude that the latter is more discriminative. This fact is due to the small number of examples of MIDDLE_NAME values within the PERSON dataset. Considering the occurrences of one attribute independently of another we may use the usual rule of thumb for Bernoulli

Distribution. That is, it is highly likely that the perplexity of FIRST_NAME, LAST_NAME, ACTIVITY, AFFILIATION, ROLE and PROVENANCE will not change with the addition of new examples, as the actual numbers are high.

We can estimate the probability that two entities selected from different documents co-refer. Actually, this is the estimate of the probability that two entities co-refer conditioned by the fact that they have been correctly identified inside the documents. We can compute such probability as the complementary of the ratio between the number of different entities and the number of the document entities in the collection.

$$P(co-ref) = 1 - \frac{\#collection-entities}{\#document-entities}$$

From Table 2 we read these values as 2574 and 3284 respectively, therefore, for the PERSON data set, the probability of intra-document co-reference is approximately 22%. We consider that this figure is only partially indicative, and that it is very likely for it to be increased after inspection of bigger corpora. This is an aposteori probability because the number of collection-entities is known only after the whole set of mentions has been processed.

An global estimator of the difficulty of the co-reference is the expectation that a correct identified mention refers to a new entity. This estimator shows the density of collection-entities in the mentions space: let us call it *co-reference density*. We can estimate the co-reference-density as the ratio between the number of different entities and the number of mentions.

$$coref-density = \frac{\#collection-entities}{\#mentions}$$

The co-reference density takes values in the interval with limits [0-1]. The case when the co-reference density tends to 0 means that all the mentions refer to the same entity, while when the value tends to 1 it means that each mention in the collection refers to a different entity. Both the limits render the co-reference task superfluous. The figure for co-reference density we found in our corpus is 2574/7233 ≈ 0.35, and it is far from being close to one of the extremes.

A measure, that can be used as a baseline for the co-reference task, is the value of co-reference density conditioned by the fact that one knows in advance whether two mentions that are identical also co-refer. Let us call this measure *pseudo-co-reference-density*. It shows the maximum accuracy of a system that deals with ambiguity by ignoring it. We approximate it as the ratio between the number of different entities and the number of distinct mentions.

$$p-coref-density = \frac{\#collection-entities}{\#distinct-mentions}$$

The pseudo-co-reference for our dataset is 2574/4851 ≈ 0.55. This information is not directly expressed in the collection, so it should be approximated. The difference between co-reference density and pseudo co-reference density shows the increase in recall, if one considers that two identical mentions refer to the same entity with probability 1. On the other hand, the loss in accuracy might be too large (consider for example the case when two different persons happen to have the same first name).

For our dataset the co-ref is ≈0,22 which means that 22% of the document entities occur in more than one document. The detailed distribution is presented in Table 11, where on the first and third columns we list the number of collection entities that occur in the number of documents that is specified in the second and fourth respectively.

#documents	#entities	#documents	#entities
1	2155 (84%)	6	6
2	286 (11%)	7	3
3	71 (2%)	8	4
4	31 (1%)	9	1
5	15 (0,5%)	16	1

Table 11. Intra-document co-reference

VII. CONCLUSION

We have presented the results of a pilot study on Ontology Population restricted to PERSON entities. One of the main motivation of the study was to individuate critical factors that determine the difficulty of the task.

The first conclusion we draw is that textual mentions of PERSON entities are highly structured. As a matter of fact, most of the mentions bring information that can be easily classified in a limited number of attributes, while only 3% of them are categorized as MISCELLANEA. These figures highly suggested that the Ontology Population from Textual Mentions (OPTM) approach is feasible and promising.

Secondly, we show that 50% of the mentions carry more than the value of a single attribute. This fact, combined with the relatively low perplexity figures for some attributes, most notably LAST_NAME, suggests a co-reference procedure based on attributes values.

Thirdly, we have computed the values of three estimators of difficulty for entity co-reference. One of them, the pseudo-co-reference-density, might be naturally used as a baseline for the task. It has been also discovered that the co-reference-density is far away from its possible extremes, 0 and 1, showing that simple string matching procedures might not achieve good results.

Our future work will be focused on two main issues: (i) the use of the PERSON dataset as training corpus for resolving the entity co-reference task, as a first step towards implementing a full OPTM system; and (ii) a controlled extension of the dataset with new data in order to understand which figures are likely to remain stable.

REFERENCES

1. Almuhareb, A., Poesio, M. (2004). Attribute-based and value-based clustering: An evaluation. In: Proceedings of EMNLP 2004, Barcelona, 2004, 158-165.
2. Avancini, H., Lavelli, A., Magnini, B., Sebastiani, F., Zanoli, R. (2003). Expanding Domain-Specific Lexicons by Term Categorization. In: Proceedings of SAC 2003, 793-79.
3. Bontcheva, K., Cunningham, H. (2003). The Semantic Web: A New Opportunity and Challenge for HLT. In: Proceedings of the Workshop HLT for the Semantic Web and Web Services at ISWC 2003, Sanibel Island, 2003.
4. Bouquet, P., Serafini, L., and Zanobini S.. Semantic coordination: a new approach and an application, In Sencond Internatinal Semantic Web Conference, volume 2870 of Lecture Notes in Computer Science, pages 130--145. Springer Verlag, September 2003
5. Buitelaar P., Cimiano P. and Magnini B. (Eds.) Ontology Learning from Text: Methods, Evaluation and applications. IOS Press, 2005.
6. Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson G. (2005). TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, MITRE.
7. Lavelli, A., Magnini, B., Negri, M., Pianta, E., Speranza, M., Sprugnoli, R. (2005). Italian Content Annotation Bank (I-CAB): Temporal Expressions (V. 1.0.). Technical Report T-0505-12. ITC-irst, Trento.
8. Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In: Proceedings of COLING-ACL98, Montreal, Canada, 1998.
9. Linguistic Data Consortium (2004). ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, version 5.6.1 2005.05.23.
10. B. Magnini, E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, V. Bartalesi Lenzi and R. Sprugnoli. I-CAB: the Italian Content Annotation Bank, In: Proceedings of LREC-2006, Genova, Italy.
11. B. Magnini, E. Pianta, O. Popescu and M. Speranza. Ontology Population from Textual Mentions: Task Definition and Benchmark. Proceedings of the OLP2 workshop on Ontology Population and Learning, Sidney, Australia, 2006. Joint with ACL/Coling 2006.
12. Tanev H. and Magnini B. Weakly Supervised Approaches for Ontology Population. Proceedings of EACL-2006, Trento, 3-7 April, 2006.
13. Velardi, P., Navigli, R., Cuchiarelli, A., Neri, F. (2004). Evaluation of Ontolearn, a Methodology for Automatic Population of Domain Ontologies. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.): Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, 2005.