Ontology Population from Textual Mentions: Task Definition and Benchmark

Bernardo Magnini, Emanuele Pianta, Octavian Popescu and Manuela Speranza

ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica
Via Sommarive 18, 38050 Povo (TN), Italy
{magnini, pianta, popescu, manspera}@itc.it

Abstract

In this paper we propose and investigate Ontology Population from Textual Mentions (OPTM), a sub-task of Ontology Population from text where we assume that mentions for several kinds of entities (e.g. PERSON, ORGANIZATION, LOCATION, GEO-POLITICAL_ ENTITY) are already extracted from a document collection. On the one hand, OPTM simplifies the general Ontology Population task, limiting the input textual material; on the other hand, it introduces challenging extensions to Ontology Population restricted to named entities, being open to a wider spectrum of linguistic phenomena. We describe a manually created benchmark for OPTM and discuss several factors which determine the difficulty of the task.

1 Introduction

Mentions are portions of text which refer to entities[1]. As an example, given a particular textual context, both the 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 "*U.S. President*".

In this paper we propose and investigate Ontology Population from Textual Mentions (OPTM), a sub-task of Ontology Learning and Population (OLP) from text where we assume that mentions for several kinds of entities (e.g. PERSON, ORGANIZATION, LOCATION, GEO-POLITICAL _ENTITY) are already extracted from a document collection.

We assume an ontology with a set of classes $C=\{c_1, ..., c_n\}$ with each class c_1 being described by a set of attribute value pairs $[a_1, v_1]$. Given a set of mentions $M=\{m_{1,c1}, ..., m_{n,cn}\}$, where each mention m_j is classified into a class c_i in C, the OPTM task is defined in three steps: Recognition and Classification of Entity Attributes, Normalization, and Resolution of inter-text Entity Coreference.

(i) **Recognition and Classification of Entity Attributes (RCEA)**. The textual material expressed in a mention is extracted and distributed along the attribute-value pairs already defined for the class c_i of the mention; as an example, given the PERSON mention "*U.S. President Bush*", we expect that the attribute LAST_NAME is filled with the value "*Bush*" and the attribute ROLE is filled with the value "*U.S. President*". Note that fillers, at this step, are still portions of text.

(ii) **Normalization**. The textual material extracted at step (i) is assigned to concepts and relations already defined in the ontology; for example, the entity BUSH is created as an instance of COUNTRY_PRESIDENT, and an instance of the relation PRESIDENT_OF is created between BUSH and U.S.A. At this step different instances are created for co-referring mentions.

(iii) **Resolution of inter-text Entity Coreference (REC)**. Each mention m_j has to be assigned to a single individual entity belonging to a class in C. For example, we recognize that the instances created at step (i) for "*U.S. President Bush*" and "*George W. Bush*" actually refer to the same entity.

[1] The terms "mention" and "entity" have been introduced within the ACE Program (Linguistic Data Consortium, 2004). "Mentions" are equivalent to "referring expressions" and "entities" are equivalent to "referents", as widely used in computational linguistics. In this paper, we use italics for "*mentions*" and small caps for ENTITY and ENTITY_ATTRIBUTE.

In this paper we address steps (i) and (iii), while step (ii) is work in progress. The input of the OPTM task consists of classified mentions and the output consists of individual entities filled with textual material (i.e. there is no normalization) with their co-reference relations. The focus is on the definition of the task and on an empirical analysis of the aspects that determine its complexity, rather than on approaches and methods for the automatic solution of OPTM.

There are several advantages of OPTM which make it appealing for OLP. First, mentions provide an obvious simplification with respect to the more general Ontology Population from text (cf. Buitelaar et al. 2005); in particular, mentions are well defined and there are systems for automatic mention recognition. Although there is no univocally accepted definition for the OP task, a useful approximation has been suggested by (Bontcheva and Cunningham, 2005) as Ontology Driven Information Extraction with the goal of extracting and classifying instances of concepts and relations defined in a Ontology, in place of filling a template. A similar task has been approached in a variety of perspectives, including term clustering (Lin, 1998 and 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). However, 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 difficulty, which makes OPTM challenging both for the Computational Linguistic side and the Knowledge Representation community. Second, there already exist annotated data with mentions, delivered under the ACE (Automatic Content Extraction) initiative (Ferro et al. 2005, Linguistic Data Consortium 2004), which makes the exploitation of machine learning based approaches possible. Finally, having a limited scope with respect to OLP, the OPTM task allows for a better estimation of performance; in particular, it is possible to evaluate more easily 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.

In the paper we both define the OPTM task and describe an OPTM benchmark, i.e. a document collection annotated with mentions as well as an ontology where information from mentions has been manually extracted. The general architecture of the OPTM task has been sketched above, considering three sub tasks. The document collection we use consists of about 500 Italian news items. Currently, mentions referring to PERSON, ORGANIZATION and GEO-POLITICAL_ ENTITY have been annotated and co-references among such mentions have been established. As for the RCEA sub task, we have considered mentions referring to PERSON and have built a knowledge base of instances, each described with a number of attribute-value pairs.

The paper is structured as follows. Section 2 provides the useful background as far as mentions and entities are concerned. Section 3 defines the OPTM task and introduces the dataset we have used, as well as the annotation procedures and guidelines we have defined for the realization of the OPTM benchmark corpus. Section 4 reports on a number of quantitative and qualitative analyses of the OPTM benchmark aimed at determining the difficulty of the task. Finally, Section 5 proposes future extensions and developments of our work.

2 Mentions and Entities

As indicated in the ACE Entity Detection task, the annotation of entities (e.g. PERSON, ORGANIZATION, LOCATION and GEO-POLITICAL_ENTITY) requires that the entities mentioned in a text be detected, their syntactic head marked, their sense disambiguated, and that selected attributes of these entities be extracted and merged into a unified representation for each entity.

As it often happens that the same entity is mentioned more than once in the same text, two inter-connected levels of annotation have been defined: 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. For instance, if the entity GEORGE_W._BUSH (e.g. the individual in the world who is the current president of the U.S.) is mentioned in two different sentences of a text as "*the U.S. president*" and as "*the president*", these two expressions are considered as two co-referring entity mentions.

The kinds of reference made by entities to something in the world are described by the following four classes:

- **specific referential entities** are those where the entity being referred to is a unique object

or set of objects (e.g. "*The president of thecompany* is here*"*)[2];

- **generic referential entities** refer to a kind or type of entity and not to a particular object (or set of objects) in the world (e.g. "*The president* is elected every 5 years");
- **under-specified referential entities** are non-generic non-specific references, including imprecise quantifications (e.g. "*everyone*") and estimates (e.g. "*more than 10.000 people*");
- **negatively quantified entities** refer to the empty set of the mentioned type of object (e.g. "*No lawyer*").

The textual extent of mentions is defined as the entire nominal phrase used to refer to an entity, thus including modifiers (e.g. "*a big family*"), prepositional phrases (e.g. "*the President of the Republic*") and dependent clauses (e.g. "*the girl who is working in the garden*").

The classification of entity mentions is based on syntactic features; among the most significant categories defined by LDD (Linguistic Data Consortium 2004) there are:

- NAM: proper names (e.g. "*Ciampi*", "*the UN*");
- NOM: nominal constructions (e.g. "*good children*", "*the company*");
- PRO: pronouns, e.g. personal ("*you*") and indefinite ("*someone*");
- WHQ: wh-words, such as relatives and interrogatives (e.g. "*Who's there?*");
- PTV: partitive constructions (e.g. "*some of them*", "*one of the schools*");
- APP: appositive constructions (e.g. "*Dante, famous poet*", "*Juventus, Italian football club*")*.*

Since the dataset presented in this paper has been developed for Italian, some new types of mentions have been added to those listed in the LDC guidelines; for instance, we have created a specific tag, ENCLIT, to annotate the clitics whose extension can not be identified at word-level (e.g. "*veder*[*lo*]"/"to see *him*"). Some types of mentions, on the other hand, have been eliminated; this is the case for pre-modifiers, due to syntactic differences between English, where both adjectives and nouns can be used as pre-modifiers, and Italian, which only admits adjectives in that position.

In extending the annotation guidelines, we have decided to annotate all conjunctions of entities, not only those which share the same modifiers as indicated in the ACE guidelines, and to mark them using a specific new tag, CONJ (e.g. "*mother and child*")[3].

According to the ACE standards, each distinct person or set of people mentioned in a document refers to an entity of type PERSON. For example, people may be specified by name ("*John Smith*"), occupation ("*the butcher*"), family relation ("*dad*"), pronoun ("*he*"), etc., or by some combination of these.

PERSON (PE), the class we have considered for the Ontology Population from Textual Mention task, is further classified with the following subtypes:

- INDIVIDUAL_PERSON: PEs which refer to a single person (e.g. "*George W. Bush"*);
- GROUP_PERSON: PEs which refer to more than one person (e.g. "*my parents", "your family"*, etc.);
- INDEFINITE_PERSON: a PE is classified as indefinite when it is not possible to judge from the context whether it refers to one or more persons (e.g. "I wonder *who* came to see me*"*).

3 Task definition

In Section 3.1 we first describe the document collection we have used for the creation of the OPTM benchmark. Then, Section 3.2 provides details about RCEA, the first step in OPTM.

3.1 Document collection

The OPTM benchmark is built on top of a document collection (I-CAB, Italian Content Annotated Bank)[4] annotated with entity mentions. I-CAB (Magnini et al. 2006) consists of 525 news documents taken from the local newspaper 'L'Adige'[5]. The selected news stories belong to four different days (September, 7th and 8th 2004 and October, 7th and 8th 2004) and are grouped into five categories: News Stories, Cultural News, Economic News, Sports News and Local News (see Table 1).

	09/07	09/08	10/07	10/08	Total
News	23	25	18	21	87
Culture	20	18	16	18	72
Economy	13	15	12	14	54
Sport	29	41	27	26	123
Local	46	43	49	51	189
TOTAL	131	142	122	130	525

Table 1: Number of news stories per category.

[2] Notice that the corpus is in Italian, but we present English examples for the sake of readability.

[3] Appositive and conjoined mentions are complex constructions. Although LDC does not identify heads for complex constructions, we have decided to annotate all the extent as head.

[4] A demo is available at http://ontotext.itc.it/webicab

[5] http://www.ladige.it/

I-CAB is further divided into training and test sections, which contain 335 and 190 documents respectively. In total, I-CAB consists of around 182,500 words: 113,500 and 69,000 words in the training and the test sections respectively (the average length of a news story is around 339 words in the training section and 363 words in the test section).

The annotation of I-CAB is being carried out manually, as we intend I-CAB to become a benchmark for various automatic Information Extraction tasks, including recognition and normalization of temporal expressions, entities, and relations between entities (e.g. the relation affiliation connecting a person to the organization to which he or she is affiliated).

3.2 Recognition and Classification

As stated in Section 1, we assume that for each type of entity there is a set of attribute-value pairs, which typically are used for mentioning that entity type. The same entity may have different values for the same attribute and, at this point no normalization of the data is made, so there is no way to differentiate between different values of the same attribute, e.g. there is no stipulation regarding the relationship between "*politician*" and "*political leader*". Finally, we currently assume a totally flat structure among the possible values for the attributes.

The work we describe in this Section and in the next one concerns a pilot study on entities of type PERSON. After an empirical investigation on the dataset described in Section 3.1 we have assumed that the attributes listed in the first column of Table 2 constitute a proper set for this type of entity. The second column lists some possible values for each attribute.

The textual extent of a value is defined as the maximal extent containing pertinent information. For instance, if we have a person mentioned as "*the thirty-year-old sport journalist*", we will select "*sport journalist*" as value for the attribute ACTIVITY. In fact, the age of the journalist in not pertinent to the activity attribute and is left out, whereas "*sport*" contributes to specifying the activity performed.

As there are always less paradigmatic values for a given attribute, we shortly present further the guidelines in making a decision in those cases. Generally, articles and prepositions are not admitted at the beginning of the textual extent of a value, an exception being made in the case of articles in nicknames.

Attributes	Possible values
FIRST_NAME	*Ralph, Greg*
MIDDLE_NAME	*J., W.*
LAST_NAME	*McCarthy, Newton*
NICKNAME	*Spider, Enigmista*
TITLE	*prof., Mr.*
SEX	*actress*
ACTIVITY AFFILIATION ROLE	*journalist, doctor* *The New York Times* *director, president*
PROVENIENCE	*South American*
FAMILY_RELATION	*father, cousin*
AGE_CATEGORY	*boy, girl*
MISCELLANEA	*The men with red shoes*

Table 2. Attributes for PERSON.

Typical examples for the TITLE attribute are "*Mister*", "*Miss*", "*Professor*", etc. We consider as TITLE the words which are used to address people with special status, but which do not refer specifically to their activity. In Italian, professions are often used to address people (e.g. "*avvocato/lawyer*", "*ingegnere/engineer*"). In order to avoid a possible overlapping between the TITLE attribute and the ACTIVITY attribute, professions are considered values for title only if they appear in abbreviated forms ("*avv.*", "*ing.*" etc.) before a proper name.

With respect to the SEX attribute, we consider as values all the portions of text carrying this information. In most cases, first and middle names are relevant. In addition, the values of the SEX attribute can be gendered words (e.g. "*Mister*" vs. "*Mrs.*", "*husband*" vs. "*wife*") and words from grammatical categories carrying information about gender (e.g. adjectives).

The attributes ACTIVITY, ROLE, AFFILIATION are three strictly connected attributes. ACTIVITY refers to the actual activity performed by a person, while ROLE refers to the position they occupy. So, for instance, "*politician*" is a possible value for ACTIVITY, while "*leader of the Labour Party*" refers to a ROLE. 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. Consider the following three mentions of the same entity:

(1) "*the journalist of Radio Liberty*"
(2) "*the redactor of breaking news*"
(3) "*a spare time astronomer*"

These three mentions lead to three different groups of ACTIVITY, ROLE and AFFILIATION. The obvious inference that the first two mentions conceptually belong to the same group is not drawn. This step is to be taken at a further stage.

The PROVENIENCE attribute can have as values all phrases denoting geographical/racial origin or provenience and religious affiliation.
The attribute AGE_CATEGORY can have either numerical values, such as "*three years old*", or words indicating age, such as "*middle-aged*", etc.
In the next section we will analyze the occurrences of the values of these attributes in a news corpus.

4 Data analysis

The difficulty of the OPTM task is directly correlated to four factors: (i) the extent to which the linguistic form of mentions varies; (ii) the perplexity of the values of the attributes; (iii) the size of the set of the potential co-references and (iv) the number of different mentions per entity. In this section we present the work we have undertaken so far and the results we have obtained regarding the above four factors.

We started with a set of 175 documents belonging to the I-CAB corpus (see Section 3.1). Each document has been manually annotated observing the specifications described in Section 3.2. We focused on mentions referring to INDIVIDUAL PERSON (Mentions in Table 3), excluding from the dataset both mentions referring to different entity types (e.g. ORGANIZATION) and PERSON GROUP. In addition, for the purposes of this work we decided to filter out the following mentions: (i) mentions consisting of a single pronoun; (ii) nested mentions, (in particular in the case where a larger mention, e.g. "*President Ciampi*", contained a smaller one, e.g. "*Ciampi*", only the larger mention was considered). The total number of remaining mentions (Meaningful mentions in Table 3) is 2343. Finally, we filtered out repetitions of mentions (i.e. string equal) that co-refer inside the same document, obtaining a set of 1139 distinct mentions.

The average number of mentions for an entity in a document is 2.09, while the mentions/entity proportion within the whole collection is 2.68. The detailed distribution of mentions with respect to document entities is presented in Table 4. Columns 1 and 3 list the number of mentions and columns 2 and 4 list the number of entities which are mentioned for the respective number of times (from 1 to 9 and more than 10). For instance, in the dataset there are 741 entities which, within a single document, have just one mention, while there are 27 entities which are mentioned more than 10 times in the same document. As an indication of variability, only 14% of document entities have been mentioned in two different ways.

Documents	175
Words	57 033
Words in mentions	8116
Mentions	3157
Meaningful mentions	2343
Distinct mentions	1139
Document entities	1117
Collection entities	873

Table 3. Documents, mentions and entities in the OPTM dataset.

#M/E	#occ	#M/E	#occ
1	741	6	15
2	164	7	11
3	64	8	12
4	47	9	5
5	31	≥10	27

Table 4. Distribution of mentions per entity.

4.1 Co-reference density

We can estimate the a priori 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 complement of the ratio between the number of different entities and the number of the document entities in the collection.

$$P(cross-coref) = 1 - \frac{\#collection-entities}{\#document-entities}$$

From Table 3 we read these values as 873 and 1117 respectively, therefore, for this corpus, the probability of intra-document co-reference is approximately 0.22.

A cumulative factor in estimating the difficulty of the co-reference task is the ratio between the number of different entities and the number of mentions. We call this ratio the *co-reference density* and it shows the a priori expectation that a correct identified mention refers to a new entity.

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

The co-reference density takes values in the interval with limits [0-1]. The case where the co-reference density tends to 0 means that all the mentions refer to the same entity, while where the value tends to 1 it means that each mention in the collection refers to a different entity. Both limits render the co-reference task superfluous. The figure for co-reference density we found in our corpus is 873/2343 ≈ 0.37, and it is far from being close to one of the extremes.

A last measure we introduce is the ratio between the number of different entities and the number of distinct mentions. Let's call it *pseudo co-reference density*. In fact it shows the value of co-reference density conditioned by the fact that one knows in advance whether two mentions that are identical also co-refer.

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

The pseudo co-reference for our corpus is 873/1139 ≈ 0.76. 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 (see Table 5) 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 people happen to have the same first name).

co-reference density	0.37
pseudo co-reference density	0.76
cross co-reference	0.22

Table 5. A priori estimation of difficulty of co-reference

4.2 Attribute variability

The estimation of the variability of the values for a certain attribute is given in Table 6. The first column indicates the attribute under consideration; the second column lists the total number of mentions of the attribute found in the corpus; the third column lists the number of different values that the attribute actually takes and, between parentheses, its proportion over the total number of values; the fourth column indicates the proportion of the occurrences of the attribute with respect to the total number of mentions (distinct mentions are considered).

Attributes	**total occ.**	**distinct occ. (%)**	**occ. prob.**
FIRST_NAME	535	303 (44%)	27,0%
MIDDLE_NAME	25	25 (100%)	2,1%
LAST_NAME	772	690 (11%)	61,0%
NICKNAME	14	14 (100%)	1,2%
TITLE	12	10 (17%)	0,8%
SEX	795	573 (23%)	51,0%
ACTIVITY	145	88 (40%)	7,0%
AFFILIATION	134	121 (10%)	11,0%
ROLE	155	92 (42%)	8,0%
PROVENIENCE	120	80 (34%)	7,3%
FAMILY_REL.	17	17(100%)	1,4%
AGE_CATEGORY	31	31(100%)	2,7%
MISCELLANEA	106	106 (100%)	9,3%

Table 6. Variability of values for attributes.

In Table 7 we show the distribution of the attributes inside one mention. That is, we calculate how many times one entity contains more than one attribute. Columns 1 and 3 list the number of attributes found in a mention, and columns 2 and 4 list the number of mentions that actually contain that number of values for attributes.

#attributes	#mentions	#attributes	#mentions
1	398	5	55
2	220	6	25
3	312	7	8
4	117	8	4

Table 7. Number of attributes inside a mention.

An example of a mention from our dataset that includes values for eight attributes is the following:

The correspondent of Al Jazira, Amr Abdel Hamid, an Egyptian of Russian nationality…

We conclude this section with a statistic regarding the coverage of attributes (miscellanea excluded). There are 7275 words used in 1139

distinct mentions, out of which 3606, approximately 49%, are included in the values of the attributes.

5 Conclusion and future work

We have presented work in progress aiming at a better definition of the general OLP task. In particular we have introduced Ontology Population from Textual Mentions (OPTM) as a simplification of OLP, where the source textual material are already classified mentions of entities.
An analysis of the data has been conducted over a OPTM benchmark manually built from a corpus of Italian news. As a result a number of indicators have been extracted that suggest the complexity of the task for systems aiming at automatic resolution of OPTM.

Our future work is related to the definition and extension of the OPTM benchmark for the normalization step (see Introduction). For this step it is crucial the construction and use of a large-scale ontology, including the concepts and relations referred by mentions. A number of interesting relations between mentions and ontology are likely to emerge.

The work presented in this paper is part of the ONTOTEXT project, a larger initiative aimed at developing text mining technologies to be exploited in the perspective of the Semantic Web. The project focuses on the study and development of innovative knowledge extraction techniques for producing new or less noisy information to be made available to the Semantic Web. ONTOTEXT addresses three key research aspects: annotating documents with semantic and relational information, providing an adequate degree of interoperability of such relational information, and updating and extending the ontologies used for Semantic Web annotation. The concrete evaluation scenario in which algorithms will be tested with a number of large-scale experiments is the automatic acquisition of information about people from newspaper articles.

6 Acknowledgements

This work was partially funded the three-year project ONTOTEXT[6] funded by the Provincia Autonoma di Trento. We would like to thank Nicola Tovazzi for his contribution to the annotation of the dataset.

[6] http://tcc.itc.it/projects/ontotext

References

Almuhareb, A. and Poesio, M.. 2004. Attribute-based and value-based clustering: An evaluation. In Proceedings of EMNLP 2004, pages 158--165, Barcelona, Spain.

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.

Cunningham, H. and Bontcheva, K. Knowledge Management and Human Language: Crossing the Chasm. Journal of Knowledge Management, 9(5), 2005.

Buitelaar, P., Cimiano, P. and Magnini, B. (Eds.) Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, 2005.

Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson, G. (2005). TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, MITRE.

Lavelli, A., Magnini, B., Negri, M., Pianta, E., Speranza, M. and Sprugnoli, R. (2005). Italian Content Annotation Bank (I-CAB): Temporal Expressions (V. 1.0.). Technical Report T-0505-12. ITC-irst, Trento.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In: Proceedings of COLING-ACL98, Montreal, Canada, 1998.

Linguistic Data Consortium (2004). ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, version 5.6.1 2005.05.23. http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v5.6.1.pdf

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V. and Sprugnoli, R. (2006). I-CAB: the Italian Content Annotation Bank. Proceedings of LREC-2006, Genova, Italy, 22-28 May, 2006.

Tanev, H. and Magnini, B. Weakly Supervised Approaches for Ontology Population. Proceedings of EACL-2006, Trento, Italy, 3-7 April, 2006.

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.