Two Sides of a Coin
Translate while Classify Multilanguage Annotations with Domain Ontology-driven Word Sense Disambiguation

Massimiliano Gioseffi and Angela Locoro
DIBRIS, Sede di Valle Puggia, Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi, University of Genova, Genova, Italy

Keywords: Domain-driven Word Sense Disambiguation, Domain Ontologies, Text Classification, Natural Language Processing, Machine Translation.

Abstract: In this paper we present an approach for the translation and classification of short texts in one step. Our work lays in the tradition of Domain-Driven Word Sense Disambiguation, though a major emphasis is given to domain ontologies as the right tool for sense-tagging and topic detection of short texts which, by their nature, are known to be reluctant to statistical treatment. We claim that in a scenario where users can annotate knowledge items using different languages, domain ontologies can prove very suitable for driving the word disambiguation and topic classification tasks. In this way, two tasks are gainfully collapsed in a single one. Although this study is still in its infancy, in what follows we are able to articulate motivations, design, workflow analysis, and concrete evolutions envisioned for our tool.

1 INTRODUCTION AND MOTIVATION

Whenever a Web user enters the Google Image (Generalistic Search Engine) Area[1] and looks for pictures that have been annotated with short captions like

"a *player* reading the *score*"

at least three different subject types are shown as top 10 ranked results: a music player, a football player and a videogame player. The words *player* and *score* have two different meanings in this case, depending on the domain in which they are exploited. The first two glosses of the words *player* and *score* respectively, as they have been taken from WordNet Online[2], are reported below:

player

1. a person who participates in or is skilled at some game;
2. someone who plays a musical instrument.

score

1. a number or letter indicating quality performance;
2. a written form of a musical composition.

Scenarios demanding automatic or semi-automatic services for searching Web contents, translating their annotations while classifying them according to their topic are more and more reaching the surface of the user's needs iceberg. Users clamour for a domain-oriented systematisation of available online information with the less effort and the more effectiveness.

A Web user trying to collect pictures of famous musicians or a philharmonic institution engaged in the enrichment of its local repository with Web contents, or a music community wanting to exchange domain digital artifacts with worldwide experts are all examples of subjects interested in services that should be able to provide a selection, a translation and a classification of Web contents based on their topic of interest.

In this paper we present an approach for the disambiguation of words in sentences by means of domain ontologies (i.e. semantic objects able to describe how entities relate, interact and should be in-

[1] https://www.google.it/imghp?hl=en&tab=wi. Last accessed on 21st October 2012.

[2] http://wordnetweb.princeton.edu/perl/webwn. Last accessed on 21st October 2012.

Gioseffi M. and Locoro A..
Two Sides of a Coin - Translate while Classify Multilanguage Annotations with Domain Ontology-driven Word Sense Disambiguation.
DOI: 10.5220/0004330203580363
In *Proceedings of the 5th International Conference on Agents and Artificial Intelligence* (ICAART-2013), pages 358-363
ISBN: 978-989-8565-39-6
Copyright © 2013 SCITEPRESS (Science and Technology Publications, Lda.)

terpreted in a specialised piece of reality), which is able to frame the translation of short sentences into their correct context, hence providing the right sense for each word to be translated. Once all words in a phrase have been sense-tagged with ontology concepts, the domain of discourse can be extracted from it in a straightforward way. As a consequence, a classification by topic for all the sources being annotated with those short texts can be provided for free.

Although our study is still in its infancy, we believe that what follows is able to provide a worthy articulation of our approach. The paper is organised as follows: Section 2 outlines the related work on Domain Driven Word Sense Disambiguation and the main differences with our work, Section 3 presents the design and the workflow of our system, whereas Section 4 discusses the evolutions envisioned for our approach. Section 5 concludes.

2 RELATED WORK

Domain-Driven or Domain-Oriented Word Sense Disambiguation (Navigli, 2009) is strongly focused on providing the most appropriate sense label for a word that is being used in domain-specific texts. The peculiarity of this approach with respect to classical Word Sense Disambiguation, according to Navigli, lays in the paradigm "*shift from linguistic understanding to a domain-oriented type-based vision of sense ambiguity*". This is especially true for cross-lingual Word Sense Disambiguation, where the domain information of a phrase may result crucial for bringing positive chances of a close translation.

A major source of domain information for the disambiguation of words has been in recent years the WordNet lexical database, as witnessed by several research studies (Gliozzo et al., 2004), (Cucchiarelli and Velardi, 1998), (Buitelaar and Sacaleanu, 2001). In these scenarios WordNet is used as a domain semantic model, especially in its version where synsets are tagged with domain labels[3]. Based on such models, score formulas are computed to determine the predominant sense of a word in a text. However, WordNet is not a proper domain ontology. Moreover, most of these techniques rely on a trained corpus (Koeling and McCarthy, 2008) (e.g. SemCor[4] and the like) as a knowledge source, instead of a domain ontology.

Notably, a recent study (Agirre et al., 2009) enforces evidences in favour of knowledge-based methods (among which we include domain ontologies) for boosting the disambiguation task in domain-specific environments. The authors claim that, when tagging domain-specific corpora, knowledge-based Word Sense Disambiguation is performing better than generic supervised Word Sense Disambiguation systems trained on generalistic corpora. The test was conducted on 41 domain-related and highly polysemous words in the two domains of Sports and Finance. The algorithm used is called Personalised Page Rank and was applied to WordNet graph in order to rank word senses.

These researches were conducted as a monolingual task. In addition, very few attempts have been made in the direction of developing Domain-Driven Word Sense Disambiguation to real case applications.

The Omega ontology (Philpot et al., 2010) was conceived as a synthesis of WordNet and Mikrokosmos (O'Hara et al., 1998), (Mahesh, 1996), a conceptual resource properly designed to support translation. Besides the core concept base, Omega was designed to connect with a range of auxiliary knowledge sources, including domain ontologies, incorporated into the basic conceptual structure and representation.

In this paper we try to extend these directions of research by exploiting ontologies conceived by domain experts as our knowledge source, and short texts annotations of domain specific digital sources as our target of disambiguation, translation and classification tasks.

3 SYSTEM ARCHITECTURE

In this Section we will briefly depict the main steps of our approach, and will give more details of the disambiguation and classification algorithm.

3.1 System Workflow

Figure 1 shows the main components, outputs and data support sources of our system.

The purpose of our approach is the translation and classification of a sentence in English into a sentence in Italian by means of a domain ontology-driven word sense disambiguation algorithm. The classification by topic of the target sentence is obtained thanks to the ontology that has been acknowledged to represent the correct domain of both the source and the target sentences after the execution of the domain driven disambiguation procedure. The main steps of the algorithm are depicted in the sequel. For sake of clarity the sentence in English

[3]http://wndomains.fbk.eu/.

[4]http://multisemcor.fbk.eu/semcor.php.

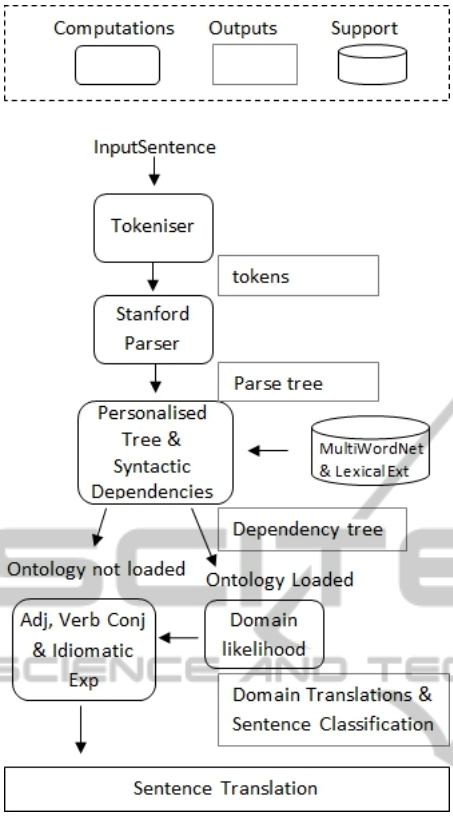

Figure 1: Workflow of a typical translation and classification session with our algorithm.

"the violin player is reading the score"

will be used as a pointwise example throughout the following description, that is also recalled in the visual workflow of Figure 1:

1. Tokenisation and lemmatisation of the sentence. The output of these two basic steps will be the phrase form:

   [the, violin, player, be, read, the, score].

2. Parsing with the Stanford Parser[5]. The output of this phase will be a Parse Tree with words tagged with their part-of-speech (POS), and so on.

3. Creation of our Program Tree based on collapsing the Stanford Parse Tree POS structure, and adding syntactic dependencies to it, as depicted in Figure 2. In the specific example the POS nodes used in the Custom Tree are:

   [the violin player]
   NP_NODE ← { NP }

   [is reading the score]
   VP_NODE ← { VP }

   [the violin], [the score]
   NOUN_NODE ← { DT, NN }

   [is reading]
   VERB_NODE ← { VBZ, VBG }

   and some of the word-pair dependencies collected for the sentence are of the kind:

   nsubj(*reading-5*, *player-3*)
   aux(*reading-5*, *is-4*)
   det(*score-7*,*the-6*)

   and so on.

4. For each word in the sentence a draft translation is tried by means of:
   - MultiWordNet[6] English-Italian alignment.
   - the data support sources, whose contents and interdependencies are depicted in Figure 3, if the word is not found in MultiWordNet.

5. For each noun and verb a disambiguation procedure is carried out by means of the ontology loaded and composed of different domains (an example of such an ontology is reported in Figure 4). The details of the disambiguation algorithm are reported in Section 3.2.

6. Conversion from an English grammar to an Italian grammar phrase structure. This procedure includes the execution of the following tasks:
   - Adjectives and verbs are correctly conjugated for number and genre, and verb tense, respectively.
   - Idiomatic expressions are correctly rendered.
   - The final translation is printed. In case the disambiguation has been carried out with the help of an ontology, the domain labels of each noun and verb of the phrase are shown, together with the more general label of the upper domain to which the domain labels belong (e.g. Music, Sports, and so on).

3.2 Disambiguation and Classification with Domain Ontologies

In case both an ontology like the one depicted in Figure 4 and a domain specific verb list are available, or

[5]http://nlp.stanford.edu/software/lex-parser.shtml.

[6]http://multiwordnet.fbk.eu/english/home.php.

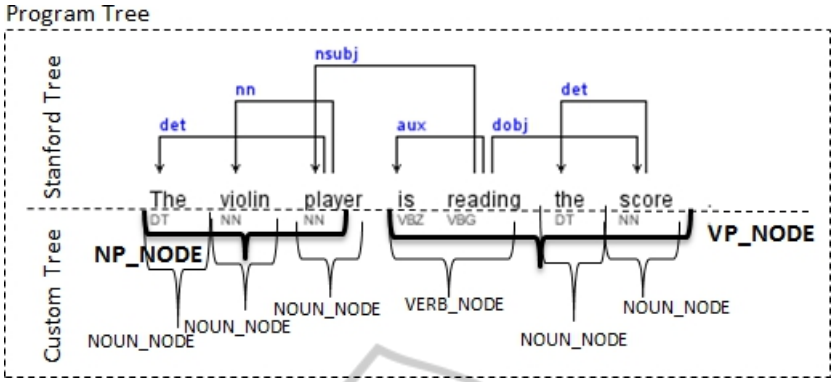

Figure 2: An example of Program Tree obtained by collapsing the POS structure of the Stanford tree with fewer compacted notations, and by extending it with word-pair dependencies.

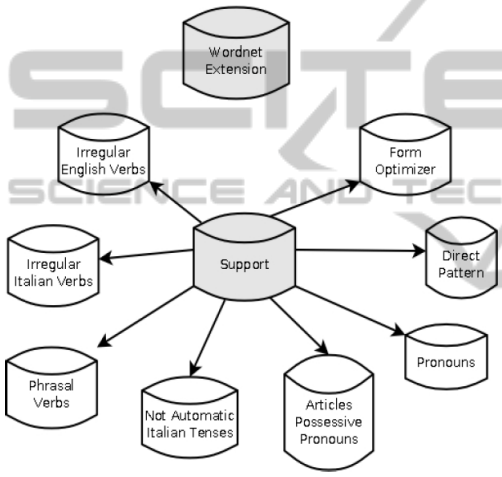

Figure 3: The data support sources for translation optimisation, seen as MultiWordNet extensions.

one of the two is available instead, the domain disambiguation of nouns (resp. verbs) of the source sentence is carried out. A specific algorithm is devoted to the computation of the likelihood for each domain and for each noun and verb in the sentence. Given w as the word to be disambiguated, t as one of its translations in the set of its possible translation T_w, c as a concept in one of the domain ontologies sub-trees O_i, v as one of the domain specific verbs of the list DV_i, and d_i as a specific domain label, the Algorithm depicted in 1 is computing the domain likelihood for each domain d_i analysed.

Starting from each sub-tree root the algorithm compares each concept of the ontology with each translation of the word being disambiguated. Each time there is an exact match between a translation of w and an ontology concept c or a verb v in the domain verbs list, the likelihood for the domain d_i is incremented by 1. The most probable domain is the one with highest likelihood (hence, with the highest number of words matching domain concepts in the ontology plus domain verbs in the verbs list, if any). The winner domain is chosen and a translation is produced according to such domain words.

Algorithm 1: DisambiguateWithOntology algorithm.

```
1:  procedure DISONT(w, T_w, O_i, DV_i)
2:      for all w ∈ Words do                ▷ noun or verbs
3:          for all t ∈ T_w do              ▷ translations of w
4:              for all c ∈ O_i, v ∈ DV_i do
5:                  if t = c, t = v then likelihood_{d_i} + 1
6:                  end if
7:              end for
8:          end for
9:      end for
10: end procedure
```

In case more than one domain results with the same likelihood score, the disambiguation is conducted with the “translation by frequency”: the top synset of MultiWordNet is taken as the ”lemma set” L_s from which the most suitable translation word $t_w \in L_s$ is selected. The selection is done by choosing the most frequent t_w in the whole space of all its synsets and glosses.

In our example, the result of the disambiguation procedure for our sentence (with domain words underlined in the English version) will be:

English: *[“the violin player is reading the score”]*
Italian: *[“il violinista sta leggendo lo spartito”]*
(Music:3,Sport:2)

and the classification results (translated in English for sake of clarity) are the following (in square brackets both the super-class of each word in the sentence, as well as the root concept of the winner domain ontology sub-tree is set):

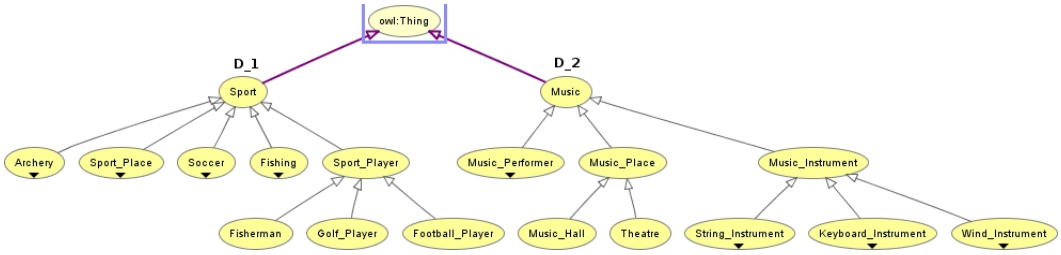

Figure 4: An ontology fragment with specific domain sub-trees under the Thing concept. The root of each sub-tree is labelled with the upper domain label (e.g. Music, Sports, and do on).

"[Domain: Music] the violin [string_instrument] player [music_performer] is reading the score [music_artefact]"

4 DISCUSSION AND NEW PROPOSALS

Our approach has potentials in the semantic treatment of texts that are by their nature short and refer to specific subjects, objects, and topics of interest, such as those that users exploit to annotate their personal or professional digital archives. Statistics is not always able to capture alone enough features when dealing with short sentences that can be found isolated from a document corpus (Wenyin et al., 2010). In addition, building a domain specific corpus or training a statistic device on an existing one may result in less slim or precise results if compared to the exploitation of a codified knowledge source as a domain ontology, which is in fact a tool especially adopted to give a structure, an organisation and a semantic description of resources in domain specific communities.

Although it is not always possible to disambiguate with a domain ontology sentences of the form:

["the player is reading the score"](Music:2,Sport:2)

as they would result in a fair likelihood for two different domains, a valid counter example could be the one where the presence of a single specialistic word may make the difference. For example:

"the strong player was playing the bass in the city orchestra near the sea and his performance was good"

may bring to both the music (with concepts: player, bass, orchestra) and the sports (with concepts: player, bass, sea) domains. However, besides the mere computation of domain words, the word *orchestra* can be considered as a "domain hapax". This phenomenon is also reflected in the position of specialistic words usually placed deep in the domain ontology hierarchy, and this can be measured, as exemplified in Table 1: the deepest the level in the ontology hierarchy, the more chances has the word (and hence, the sentence) to be assigned to that specific domain. The degree of specificity of a word could be considered as a valid criterion in the topic interpretation of a sentence, hence the ontology level reached during the disambiguation procedure can be used as its measure. In the same vein, if a set of words in a sentence belongs to more than one domain, a selection measure could be the "semantic relatedness" of such words in each domain ontology, expressed as the number of connections between pairwise concepts. The winner domain for that sentence could be the one with the highest semantic relatedness among such words, under the hypothesis that a sentence tends to express stronger relations between objects of the reality. An example is depicted in Figure 5.

Table 1: Example of the deepest level reached when visiting two domain ontology trees during the disambiguation of the above sentence. According to this measure the Music domain is chosen.

Domain ontology	Ontology level
Music	3
Sports	2

5 CONCLUSIONS AND FUTURE WORK

In this paper we have prospected an alternative direction to existing domain-driven word sense disambiguation methodologies by proposing to exploit domain ontologies. We claimed that this is a promising approach and we gave motivations to our hypothesis. Our future work will address the testing of domain ontology-driven word sense disambiguation against different translation tools, and the extension of our

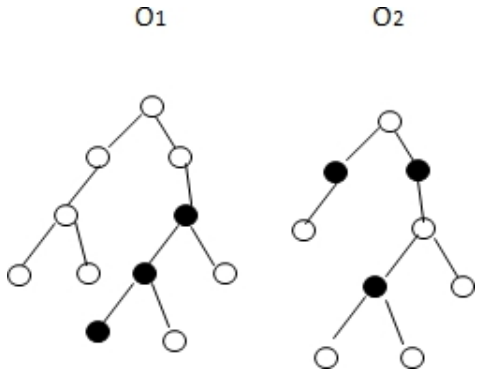

Figure 5: Semantic relatedness between (black) nodes in two domain ontologies. The more the nodes are connected (ontology O_1), the highest the chance that a sentence belongs to that domain of discourse instead of belonging to the other (ontology O_2).

knowledge source with its inclusion, for example, into the Omega ontology. We are also interested in the development of ad hoc concepts relatedness measures that can strength the hypothesis of domain disambiguation for a sentence. The application scenario envisioned for our tool is that of the translation, classification and retrieval of multilanguage annotations of digital contents.

ACKNOWLEDGEMENTS

This work is partially supported by the "Indiana MAS and the Digital Preservation of Rock Carvings: A multi-agent system for drawing and natural language understanding aimed at preserving rock carvings" MIUR FIRB Project funded by the Italian Ministry of Education, University and Research under fund identifier RBFR10PEIT.

REFERENCES

Agirre, E., De Lacalle, O. L., and Soroa, A. (2009). Knowledge-based wsd on specific domains: performing better than generic supervised wsd. In *Proceedings of the 21st international jont conference on Artifical intelligence*, IJCAI'09, pages 1501–1506, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Buitelaar, P. and Sacaleanu, B. (2001). Ranking and selecting synsets by domain relevance. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*.

Cucchiarelli, A. and Velardi, P. (1998). Finding a domain-appropriate sense inventory for semantically tagging a corpus. *Nat. Lang. Eng.*, 4(4):325–344.

Gliozzo, A. M., Magnini, B., and Strapparava, C. (2004). Unsupervised domain relevance estimation for word sense disambiguation. In *EMNLP*, pages 380–387. ACL.

Koeling, R. and McCarthy, D. (2008). From predicting predominant senses to local context for word sense disambiguation. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 129–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mahesh, K. (1996). Ontology development for machine translation: Ideology and methodology.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

O'Hara, T., Mahesh, K., and Niremburg, S. (1998). Lexical acquisition with wordnet and the mikrokosmos ontology. In *In Proceedings of the ACL Workshop on the Use of WordNet in NLP*, pages 94–101.

Philpot, A., Hovy, E., and Pantel, P. (2010). *The Omega ontology*. Cambridge University Press.

Wenyin, L., Quan, X., Feng, M., and Qiu, B. (2010). A short text modeling method combining semantic and statistical information. *Information Sciences*, 180(20):4031 – 4041.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS