# Bridging Languages by SuperSense Entity Tagging

**Davide Picca** and **Alfio Massimiliano Gliozzo*** and **Simone Campora****
University of Lausanne, CH 1015-Lausanne-Switzerland
*Semantic Technology Lab (STLab - ISTC - CNR), Via Nomentana 56-0016, Rome, Italy
**Ecole Polytechnique Federale de Lausanne (EPFL)
davide.picca@unil.ch, alfio.gliozzo@istc.cnr.it, simone.campora@gmail.com

## Abstract

This paper explores a very basic linguistic phenomenon in multilingualism: the lexicalizations of entities are very often identical within different languages while concepts are usually lexicalized differently. Since entities are commonly referred to by proper names in natural language, we measured their distribution in the lexical overlap of the terminologies extracted from comparable corpora. Results show that the lexical overlap is mostly composed by unambiguous words, which can be regarded as anchors to bridge languages: most of terms having the same spelling refer exactly to the same entities. Thanks to this important feature of Named Entities, we developed a multilingual super sense tagging system capable to distinguish between concepts and individuals. Individuals adopted for training have been extracted both by YAGO and by a heuristic procedure. The general F1 of the English tagger is over 76%, which is in line with the state of the art on super sense tagging while augmenting the number of classes. Performances for Italian are slightly lower, while ensuring a reasonable accuracy level which is capable to show effective results for knowledge acquisition.

## 1 Introduction

The Semantic Web paradigm is often required to provide a structured view of the unstructured information expressed in texts (Buitelaar et al., 2005; Cimiano, 2006). Semantic technology requires abundance of such kind of knowledge in order to cover the web scale in almost any language. Natural Language Processing (NLP) has been adopted with the purpose of knowledge acquisition, and in particular for ontology learning and information extraction. Structured information in ontologies is often expressed by taxonomies of concepts, and then populated by instances.

Nonetheless, automatically distinguish concepts from entities in taxonomies is not an easy task, especially as far as the problem of acquiring such knowledge from texts is concerned (Zirn et al., 2008; Picca and Popescu, 2007; Miller and Hristea, 2006). First of all because such a distinction is quite vague. From a description logics perspective, that is incidently widely adopted in ontology engineering, instances are the leaves of any taxonomy as they cannot be further subcategorized and populated by other instances. For example,"Bill Clinton" is clearly an individual, since it is instance of many concepts, such as *person* or *president*, but at the same time it is a non sense describing individuals belonging to the class *Bill Clinton*.

In order to tackle this issue, we aim to provide empirical evidence to a very basic linguistic phenomenon in multilingualism, which allows the exploitation of comparable corpora for bilingual lexical acquisition. It consists on the fact that the lexicalizations of entities is very often identical within different languages while concepts are usually lexicalized differently (de Pablo et al., 2006). The existence of this phenomenon is quite intuitive and can be easily justified by considering entities as often referred to by means of ostensive acts (i.e. the act of nominating objects by indicating them), performed *in presentia* during every day life. Since entities are usually referred to using proper names in natural language, we measured their distribution in the lexical overlap of the terminologies extracted from comparable corpora in two different sample languages (i.e. Italian and English).

Named Entities are instances of particular concepts (such as person or location) and are referred



*Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009*, pages 136–142,
Suntec, Singapore, 7 August 2009. 

to by proper names. Named Entity Recognition (NER) is a basic task in NLP that has the intent of automatically recognizing Named Entities. Incidentally, NER systems can be a useful step for broad-coverage ontology engineering but they have two main limitations:

- Traditional categories (e.g., person, location, and organization) are too few and generic. It is quite evident that taxonomies require more categories than the three mentioned above.

- Even though NER systems are supposed to recognize *individuals*, very often they also returns common names and no clear distinction with *concepts* is made.

A Super Sense Tagger (SST) (Ciaramita and Johnson, 2003) is an extended NER system that uses the wider set of categories composed by the 41 most general concepts defined by WordNet. WordNet has been organized according to psycholinguistic theories on the principles governing lexical memory (Beckwith et al., 1991). Thus the broadest WordNet's categories can serve as basis for a set of categories which exhaustively covers, at least as a first approximation, all possible concepts occurring in a sentence.

The aim of this paper is to develop and explore the property of instances being lexicalized identically in different languages in order to produce a SST having the following two features:

- Make explicit distinction between instances and concepts.

- Analyze the terminology of different languages adopting a common category set.

Nevertheless, the first point demands to face with the vague distinction between concepts and individuals belonging to those concepts. So one of the main issues explored in this paper is the automatic tagging of which categories clearly have this distinction.

The paper is organized as follows. In Section 2 we describe the multilingual SST, an Italian extension of the English SST that we exploited in Section 3 to show that the lexical overlap between languages is mostly composed by unambiguous words, which can be also regarded as anchors to bridge the two languages. Most of terms having the same spelling in the two languages exactly refer to the same entities. We measured those occurrencies with respect to all different ontological types identified by our tagging device, observing that most of the overlapped terms are proper names of persons, organization, locations and artifact, while the remaining ontological types are mostly lexicalized by common nouns and have a quite empty overlap. This confirms our claim that entities of tangible types are always lexicalized by the same terms.

In Section 4 we extended the SuperSense Tagger in order to distinguish instances from individuals, while Section 5 is about evaluation. Finally Section 6 concludes the paper proposing new directions for future investigation.

## 2 Multilingual Supersense Tagging

SuperSense Tagging is the problem to identify terms in texts, assigning a "supersense" category (e.g. `person`, `act`) to their senses within their context and apply it to recognize concepts and instances in large scale textual collections of texts. An example of tagging is provided here:

Guns$_{B-noun.group}$ and$_{I-noun.group}$ Roses$_{I-noun.group}$ plays$_{B-verb.communication}$ at$_{O}$ the$_{O}$ stadium$_{B-noun.location}$

These categories are extracted from WordNet. WordNet (Fellbaum, 1998) defines 45 lexicographer's categories, also called *supersenses* (Ciaramita and Johnson, 2003). They are used by lexicographers to provide an initial broad classification for the lexicon entries [1].

Although simplistic in many ways, the supersense ontology has several attractive features for NLP purposes. First of all, concepts are easily recognizable, however very general. Secondly, the small number of classes makes the implementation of state of the art methods possible (e.g. sequence taggers) to annotate text with supersenses. Finally, similar word senses tend to be merged together reducing ambiguity. This technology has been also adopted for Ontology Learning (Picca et al., 2007), as the top level WordNet supersenses cover almost any high level ontological type of interest in ontology design. Compared to other semantic tagsets, supersenses have the advantage of being designed to cover all possible open class words. Thus, in principle there is a supersense cat-

[1] We have used the WordNet version 2.0 for all the experiments in the paper.

egory for each word, known or novel. Additionally, no distinction is made between proper and common nouns, whereas standard NER systems tends to be biased towards the former.

Following the procedure described in (Picca et al., 2008), we developed a multilingual SST working on both Italian and English languages by training the same system on MultiSemcor (Bentivogli et al., 2004), a parallel English/Italian corpus composed of 116 texts which are the translation of their corresponding English texts in SemCor. This resource has been developed by manually translating the English texts to Italian. Then, the so generated parallel corpus has been automatically aligned at the Word Level. Finally, sense labels have been automatically transferred from the English words to their Italian translations.

The sense labels adopted in the Italian part of MultiSemCor (Bentivogli et al., 2004) have been extracted by Multi WordNet [2]. It is a multilingual computational lexicon, conceived to be strictly aligned with the Princeton WordNet. The available languages are Italian, Spanish, Hebrew and Romanian. In our experiment we used the English and the Italian components. The last version of the Italian WordNet contains around 58,000 Italian word senses and 41,500 lemmas organized into 32,700 synsets aligned with WordNet English synsets. The Italian synsets are created in correspondence with the Princeton WordNet synsets whenever possible, and the semantic relations are ported from the corresponding English synsets. This implies that the synset index structure is the same for the two languages.

The full alignment between the English and the Italian WordNet is guaranteed by the fact that both resources adopts the same synset IDs to refer to concepts. This nice feature has allowed us to infer the correct super-sense for each Italian sense by simply looking at the English structure. In this way, we assign exactly the same ontological types to both Italian and English terms, thus obtaining an Italian corpus tagged by its supersenses as shown below:

$\text{I}_{O}$ $\text{Guns}_{B-noun.group}$ $\text{and}_{I-noun.group}$ $\text{Roses}_{I-noun.group}$ $\text{suonano}_{B-verb.communication}$ $\text{allo}_{O}$ $\text{stadio}_{B-noun.location}$

[2] Available at *http://multi WordNet.itc.it*.

## 3 Lexical Overlap in Comparable Corpora

Comparable corpora are collections of texts in different languages that regard similar topics (e.g. a collection of news published by press agencies in the same period). More restrictive requirements are expected for parallel corpora (i.e. corpora composed of texts which are mutual translations), while the class of the multilingual corpora (i.e. collection of texts expressed in different languages without any additional requirement) is the more general. Obviously parallel corpora are also comparable, while comparable corpora are also multilingual.

In comparable corpora, most of the individuals preserve the same spelling across different languages, while most concepts are translated differently. The analysis of the acquired terms for different ontological types shows a huge percentage of overlapped Named Entities. For our experiments, we assumed that the distinction between common names and proper names reflect as well the difference between concepts and entities in a formal ontology. Since proper names are recognized by the PoS tagger with relatively high precision, we interpreted occurrences of proper names in the acquired terminology as an evidence for detecting entities.

The Leipzig Corpora Collection (Quasthoff, 2006) presents corpora in different languages using the same format and comparable sources. The corpora are identical in format and similar in size and content. They contain randomly selected sentences in the language of the corpus. For the experiments reported in this paper, we used the Italian and the English part composed by 300,000 sentences. As shown in Figure 1 and in Figure 2, Named Entities are mostly concentrated into tangible types: Groups (organizations), Locations, Persons and Artifacts.

The results analysis is more impressive. Figure 3 shows that the lexical overlap (i.e. the subset of terms in common between English and Italian) is composed almost exclusively by entities (i.e. proper nouns). Instead if we take a look at Figure 4, we can observe that concepts are generally not shared, having an average percentage lower than 0.1%, independently of the ontological type. We can also observe the predictable result that ontological categories denoting material objects (i.e. persons, locations and groups, artifacts) still have

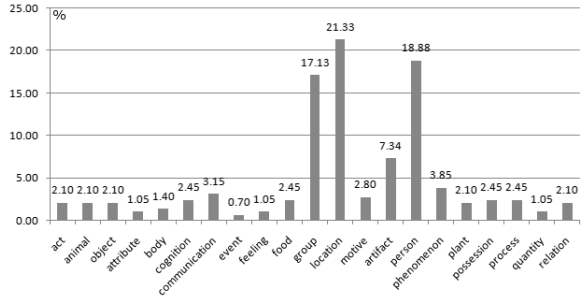


Figure 1: Distribution of discovered entity types in English

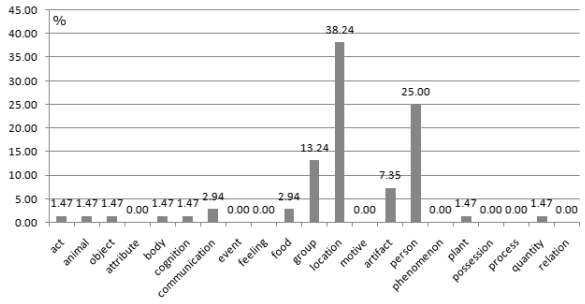


Figure 2: Distribution of discovered entity types in Italian

greater percentage of shared entities.

This is in line with the common practice of training NER on these categories. Examples of shared terms (entities) in concrete categories are:

- **noun.group**: e.g. NATO, Boeing, NASA;
- **noun.location**: e.g. Canada, Austria, Houston;
- **noun.person**: e.g. Romano_Prodi, Blair, Kofi_Annan.

Incidentally, exceptions can be found to our hypothesis (i.e. some concept is also shared).

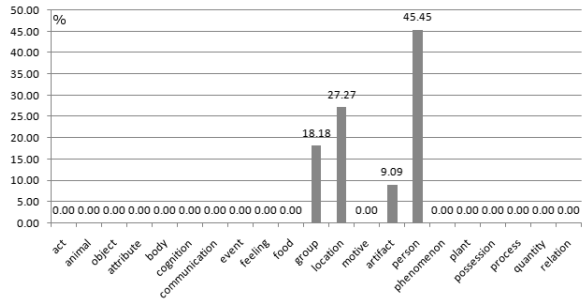


Figure 3: Shared Named Entities in both languages

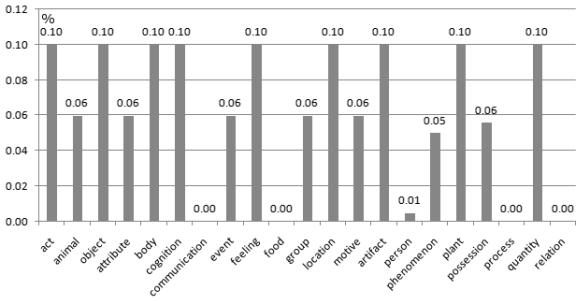


Figure 4: Shared Concepts in both languages

Examples are terms belonging to the supersense `noun.object` such as Radio and Computer. Anyhow, being them ported from one language to another, they generally do not cause problems, since they tend to share the same meaning. In our experiments (i.e. in the sample we manually analyzed), we did not find any false friend, suggesting that the impact of those words is relatively small, in spite of the fact that it is very often overemphasized.

Inversely, many abstract types (e.g. *noun.possession* and *noun.feeling*) do not share terminology at all.

## 4 Distinguishing entities from concepts

Successively, we subdivided each category into two sub-categories for both languages, *Instance* and *Concept* so that now the term "president" is tagged as *noun.person_Concept* and the term "Bill Clinton" as *noun.person_Instance*. In order to automate this task and create a reliable training set, we adopted the following strategy.

We used the concept/instances distinction provided by YAGO (Suchanek et al., 2007b). YAGO is a huge semantic knowledge base developed by the Max-Plack-Institute of Saarbrcken. YAGO knows over 1.7 million entities (like persons, organizations, cities, etc.). YAGO, exploits Wikipedia's info-boxes and category pages. Infoboxes are standardized tables that contain basic information about the entity described in the article (Suchanek et al., 2007a). For our purposes it is fundamental that YAGO's components are represented as entities. In our experiment we exploit entities as proper names and we use only YAGO entity database containing named entities.

For each term belonging to one of the concrete categories, we check if it appears in YAGO entity dataset, otherwise, if the term is not found in

YAGO, it has to satisfy all the following conditions to be tagged as *Instance*:

- The part of speech of the term belongs to one of the noun categories as "NN", "NNS", "NNP" or "NNPS".
- The first letter of the term is a capital letter.
- The term does not come after a full stop.

Upon a total of 12817 instances, almost $\frac{1}{4}$ have been found in YAGO, 3413 have been found using the heuristic strategy and the rest have been classified as concepts. If we take the previous example, the new output has now this form:

- $\text{Guns}_{B-noun.group\_Instance}$ $\text{and}_{I-noun.group\_Instance}$ $\text{Roses}_{I-noun.group\_Instance}$ $\text{plays}_{B-verb.communication}$ $\text{at}_{O}$ $\text{the}_{O}$ $\text{stadium}_{B-noun.location\_Concept}$

or

- $\text{Guns}_{B-noun.group\_Instance}$ $\text{and}_{I-noun.group\_Instance}$ $\text{Roses}_{I-noun.group\_Instance}$ $\text{suonano}_{B-verb.communication}$ $\text{allo}_{O}$ $\text{stadio}_{B-noun.location\_Concept}$

Afterwards, we trained the SST engine. It implements a Hidden Markov Model, trained with the perceptron algorithm introduced in (Collins, 2002) and it achieves a recall of 77.71% and a precision of 76.65% . Perception sequence learning provides an excellent trade-off accuracy/performance, sometimes outperforming more complex models such as Conditional Random Fields (Nguyen and Guo, 2007). We optimized the required parameters by adopting a cross validation technique. As for the settings developed by (Ciaramita and Johnson, 2003), the best results have been obtained by setting 50 trials and 10 epochs to train the perceptron algorithm. The basic feature set used for the training process, includes:

- *word* = lower-cased form of each token for the current position i and in addition for i-1 and i+1
- *sh* = shape of the token as a simple regular expression-like representation
- *pos* = POS of i, i-1 and i+1

| Category | Recall | Prec. | F1 |
|---|---|---|---|
| noun.artifact_Concept | 0.72 | 0.73 | 0.73 |
| noun.artifact_Instance | 0.59 | 0.64 | 0.62 |
| noun.group_Concept | 0.72 | 0.73 | 0.73 |
| noun.group_Instance | 0.68 | 0.70 | 0.69 |
| noun.location_Concept | 0.68 | 0.65 | 0.66 |
| noun.location_Instance | 0.75 | 0.80 | 0.77 |
| noun.person_Concept | 0.83 | 0.80 | 0.82 |
| noun.person_Instance | 0.92 | 0.88 | 0.90 |

Table 1: Recall, precision and F1 for each category for English

- *sb*= bi- and tri-grams of characters of the suffix of word_i
- *pr*= bi- and tri-grams of characters of the prefix of word_i
- *rp* = coarse relative position of word_i, rp=begin if i = 0, rp=end if i = —sentence—-1, sb=mid otherwise
- *kf* = constant features on each token for regularization purposes

Finally, we trained the SST engine in the Italian corpus generated so far, and we evaluated the super sense tagging accuracy by adopting the same evaluation method as described in (Ciaramita and Johnson, 2003), obtaining F1 close to 0.70. However quite lower than the English F1, this result is in line with the claim, since the Italian corpus is smaller and lower in quality.

## 5 SST Performance and Evaluation

We evaluated the performances of the SST generated so far by adopting a n-fold cross validation strategy on the Semcor adopted for training. Results for the chosen categories are illustrated in Table 1 and Table 2, reporting precision, recall and F1 for any Supersense. If we cast a deeper glance at the tables, we can clearly notice that for some category the F1 is exceptionally high. Some of those best categorized categories are really essential for ontology learning. For example, important labels as *noun.person* or *noun.group* achieve results among the 70%. For some categories we have found a F1 over 0.80% as *noun.person_Instance* (F1 0.90% ) or *noun.person_Concept* (F1 0.85% )

On the other hand, the Italian tagger achieved lower performances if compared with the English.

| Category | Recall | Prec. | F1 |
|---|---|---|---|
| noun.artifact_Concept | 0.64 | 0.63 | 0.63 |
| noun.artifact_Instance | 0.66 | 0.67 | 0.66 |
| noun.group_Concept | 0.61 | 0.65 | 0.63 |
| noun.group_Instance | 0.66 | 0.66 | 0.66 |
| noun.location_Concept | 0.55 | 0.53 | 0.54 |
| noun.location_Instance | 0.56 | 0.76 | 0.64 |
| noun.person_Concept | 0.81 | 0.76 | 0.78 |
| noun.person_Instance | 0.88 | 0.81 | 0.85 |

Table 2: Recall, precision and F1 for each category for Italian

It can be explained by (i) the lower quality of the training resource, (ii) the lower quantity of training data and (iii) the unavailability of the first sense info.

Regarding the first point, it is worthwhile to remark that even if the quality of transfer developed by (Bentivogli et al., 2004) is high, many incorrect sense transfers (around 14%) can be found. Because of that our work suffers of the same inherited faults by the automatic alignment. For instance, we report here the most relevant errors we faced with during the preprocessing step. One of the main errors that has badly influenced the training set especially for multiword recognition is the case in which the translation equivalent is indeed a cross-language synonym of the source expression but not a lexical unit. It occurs when a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a free combination of words (for instance *occhiali da sole* annotated with the sense of *sunglasses*).

Regarding the second problem, we noticed that the quantity of sense labeled words adopted for English is higher than 200,000, whereas the amount of Italian tokens adopted is around 92,000. Therefore, the amount of Italian training data is sensibly lower, explaining the lower performances.

Moreover, the italian SST lacks in one of the most important feature used for the English SST, first sense heuristics. In fact, for the Italian language, the first sense baseline cannot be estimated by simply looking at the first sense in WordNet, since the order of the Italian WordNet does not reflect the frequency of senses. Therefore, we did not estimate this baseline for the Italian SST, in contrast to what has been done for the English SST.

## 6 Conclusion and Future Work

In this work, we presented an empirical investigation about the role of Named Entities in comparable corpora, showing that they largely contribute in finding bridges between languages since they tend to refer to the same entities. This feature allows us to discover bridges among languages by simply looking for common Named Entities in corpora that are generally not parallels since such terms are usually associated to the same objects in the external world. We demonstrated that most terms in the lexical overlap between languages are entities, and we showed that they belong to few fundamentals categories (including persons, locations and groups).

A predominant amount of entities in the lexical overlap could be conceived as a support to our claim that Named Entities can be used to bridge the languages, since they preserve meaning and provide a set of highly accurate anchors to bridge languages in multilingual knowledge bases. Those anchors can be used as a set of seeds to boost further statistical or logical lexical acquisition processes. In addition, the impact of false friends revealed to be less problematic than expected.

We trained a multilingual super sense tagger on the Italian and English language and we introduced the distinction between concept and instance in a subset of its target classes, where our investigation suggested to look for concrete types. The resulting tagger largely extend the capabilities of the state of art supersense technology, by providing a multilingual tool which can be effectively used for multilingual knowledge induction.

For the future, we are going to further explore the direction of multilingual knowledge induction, exploiting the tagger developed so far for ontology engineering and knowledge retrieval. In addition, we plan to leverage more on the lexical overlap property analyzed in this paper, for example to develop unsupervised super sense taggers for all languages where annotated corpora are not available.

## Acknowledgments

Alfio Massimiliano Gliozzo has been supported by the BONy project, financed by the Education and culture DG of the EU, grant agreement N 135263-2007-IT-KA3-KA3MP, under the Lifelong Learning Programme 2007 managed by EACEA.

## References


R. Beckwith, C. Fellbaum, D. Gross, and G. Miller. 1991. 9. wordnet: A lexical database organized on psycholinguistic principles. *Lexicons: Using On-Line Resources to Build a Lexicon*, pages 211–232, Jan.

L. Bentivogli, P. Forner, and E. Pianta. 2004. Evaluating cross-language annotation transfer in the multisemcor corpus. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 364, Morristown, NJ, USA. Association for Computational Linguistics.

P. Buitelaar, P. Cimiano, and B. Magnini. 2005. *Ontology learning from texts: methods, evaluation and applications*. IOS Press.

M. Ciaramita and M. Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *Proceedings of EMNLP-03*, pages 168–175, Sapporo, Japan.

P. Cimiano. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

M. Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP-02*.

C. de Pablo, J.L. Martínez, and P. Martínez. 2006. Named entity processing for cross-lingual and multilingual ir applications. In *proceedings of the SIGIR2006 workshop on New Directions In Multilingual Information Access*.

C. Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. MIT Press.

G. A. Miller and F. Hristea. 2006. Wordnet nouns: Classes and instances. *Computational Linguistics*, 32(1):1–3.

N. Nguyen and Y. Guo. 2007. Comparison of sequence labeling algorithms and extensions. In *Proceedings of ICML 2007*, pages 681–688.

D. Picca and A. Popescu. 2007. Using wikipedia and supersense tagging for semi-automatic complex taxonomy construction. *In proceedings RANLP*.

D. Picca, A. Gliozzo, and M. Ciaramita. 2007. Semantic domains and supersens tagging for domain-specific ontology learning. In *proceedings RIAO 2007*.

D. Picca, A. M. Gliozzo, and M. Ciaramita. 2008. Supersense tagger for italian. In *proceedings of the sixth international conference on Language Resources and Evaluation (LREC 2008)*.

C. B. Quasthoff, U. M. Richter. 2006. Corpus portal for search in monolingual corpora,. In *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC*, pages pp. 1799–1802.

F. Suchanek, G. Kasneci, and G. Weikum. 2007a. Yago: A large ontology from wikipedia and wordnet. *Technical Report*.

F. M. Suchanek, G. Kasneci, and G. Weikum. 2007b. Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706, New York, NY, USA. ACM Press.

C. Zirn, V. Nastase, and M. Strube. 2008. Distinguishing between instances and classes in the wikipedia taxonomy. *Lecture notes in computer science*, Jan.