

# Named Entity WordNet

Antonio Toral\* Rafael Muñoz† Monica Monachini\*

\*Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Pisa, Italy  
{antonio.toral, monica.monachini}@ilc.cnr.it

†Natural Language Processing and Information Systems Group, University of Alicante, Spain  
rafael@dlsi.ua.es

## Abstract

This paper presents the automatic extension of Princeton WordNet with Named Entities (NEs). This new resource is called Named Entity WordNet. Our method maps the noun is-a hierarchy of WordNet to Wikipedia categories, identifies the NEs present in the latter and extracts different information from them such as written variants, definitions, etc. This information is inserted into a NE repository. A module that converts from this generic repository to the WordNet specific format has been developed. The paper explores different aspects of our methodology such as the treatment of polysemous terms, the identification of hyponyms within the Wikipedia categorization system, the identification of Wikipedia articles which are NEs and the design of a NE repository compliant with the LMF ISO standard. So far, this procedure enriches WordNet with 310,742 NEs and 381,043 “instance of” relations.

## 1. Introduction

Named Entity Recognition is a subtask of Information Extraction that seeks to locate and classify elements in text into a set of categories. Usually, Named Entities (NEs) refer to proper names (a.k.a. proper nouns) and numerical expressions (e.g. dates, quantities). Different NE sets of categories have been proposed, from the 4 category set (person, location, organisation, miscellaneous) of ConLL-2002 to the hierarchical proposal by Sekine with more than 100 subtypes (Sekine et al., 2002). Knowledge about NEs is important in order to tackle several Natural Language Processing (NLP) tasks. Examples include Question Answering (Mann, 2002) and Dialogue (Levy et al., 1997). Even if mature resources of geographical NEs (gazetteers) do exist (e.g. geonames<sup>1</sup>), there is a lack of more general resources. WordNet is one of the most widely used Lexical Resource (LR) within the NLP area. It is manually built by expert lexicographers and contains nouns, adjectives, verbs and adverbs organised in sets of synonyms called synsets (Miller, 1995). Several relations can be established between synsets (e.g. hyponymy, meronymy). Regarding nouns, WordNet distinguishes between common nouns (classes) and proper nouns (instances) from version 2.1. (Miller and Hristea, 2006). On the one hand, WordNet’s coverage about open domain common nouns is quite high, but on the other it contains very few proper nouns<sup>2</sup>. This is related with the following statement: “building a proper noun ontology is more difficult than building a common noun ontology as the set of proper nouns grows more rapidly” (Mann, 2002). The problem is then that a proper noun resource should be constantly updated. Therefore it is unfeasible to manually populate LRs with proper nouns. In this paper, the terms proper nouns, instances and NEs are considered synonymous and the same states for common nouns and classes. Wikipedia is an on-line multilingual encyclopedia that follows the wiki philosophy. It is built in a collaborative way by a vast number of users<sup>3</sup> and contains a huge amount of

entries<sup>4</sup>. Apart from the encyclopedic entries, Wikipedia includes interesting additional information for linguistic processing such as a category taxonomy and multilingual links.

This paper presents Named Entity WordNet (NEWN), a new resource that extends WordNet with NEs automatically extracted from the English Wikipedia. Our hypothesis is that exploiting Wikipedia is a sensible choice to automatically populate LRs with proper nouns. The main reasons are that Wikipedia is a dynamic source, contains a huge amount of proper nouns and has some degree of structure that facilitates their extraction.

The rest of the paper is organised as follows. Next section summarises related work. This is followed by the description of our method to extend WordNet with NEs. Afterwards, we present the conclusions about the approach followed and outline future work lines.

## 2. Related Work

There exist in the recent literature several publications regarding the creation of specific LRs for proper nouns (onomastica) and the enrichment of generic LRs with proper nouns.

(Sheremetyeva et al., 1998) presents the structure of a multilingual onomastica made up of a set of monolingual onomastica cross-referenced by translation links. The entries are organised in a hierarchy made up of 45 semantic categories. A semiautomatic population procedure is proposed, which is supported by an acquisition and administration interface.

Prolexbase, a multilingual database of proper names, was created within the Prolex project (Tran et al., 2004) (Krstev et al., 2005). It is based on an ontology which has four layers (instances, linguistic, conceptual and meta-conceptual) and several relations (synonymy, meronymy, antonomasia, etc). Entries are linked to EuroWordNet’s Inter-Lingual Index. The population of Prolex seems to be done manu-

<sup>1</sup>www.geonames.org

<sup>2</sup>7,669 synsets are tagged as being instances in WordNet 2.1.

<sup>3</sup>On 2007/10/29 the English version has 5,682,580 registered

users.

<sup>4</sup>On 2007/10/29 the English version contains 2,066,619 encyclopedic entries.

ally. It contains 323,000 entries and 55,000 relation links for French and 13,000 entries for German. (Mann, 2002) automatically derives a proper noun ontology from a 1 gigabyte corpus. The resulting resource contains 113,000 proper nouns and reaches 60% precision. The author remarks that it is not straight-forward to integrate the resulting resource with the WordNet noun taxonomy. (Sundheim et al., 2006) studies the linkage of a gazetteer to WordNet. The paper proposes to incorporate the instances of a geographic nature from WordNet into the Integrated Gazetteer Database (IGDB). This is justified by the fact that both resources contain complementary information. (de Loupy et al., 2004) is perhaps the most related paper to ours. It proposes to use WordNet as a proper noun thesaurus for a Question Answering system by enriching it with 130,675 proper nouns. These nouns are extracted from several knowledge bases (the authors do not specify which) and from Internet. 55 types of entries are enriched with proper nouns. However, not all of them seem to contain proper nouns (e.g. “professions” contains “Academic teacher”, “political titles” contains “1st secretary”). The methodology followed to build this thesaurus is not mentioned, which lead us to think that both the acquisition of proper nouns and their insertion in the correspondent synsets are carried out manually.

### 3. Method

A first draft of our approach has been already presented (Toral and Muñoz, 2007). Essentially it consists of two phases. In the first, WordNet’s noun synsets made up of classes<sup>5</sup> are linked to Wikipedia’s categories by matching lemmas. In the second, the entries of Wikipedia that belong to mapped categories and are NEs are incorporated into WordNet as new synsets and are linked to the input synsets with *instance of* relations.

In order to check whether an entry of Wikipedia is a NE or not, we relied on the fact that proper nouns and common nouns follow different capitalisation norms. Of course this is language dependent as not every language follows these rules. Our method collected occurrences of the extracted entries in the Web and computed the percentage of times they occur beginning by capital letters. A tuning set of 200 entries was manually tagged in order to establish an empirical threshold which was set to 91%. Thus, if an entry occurs less than 91% of the times beginning by capital letters it was classified as a class, while otherwise it was classified as a NE. This method obtained in a test set of 100 entries 93.02% precision, 68.97% recall and 83.33% F-measure $_{\beta=0.5}$ <sup>6</sup>.

This paper departs from that first proposal and explores several directions, mainly: mapping analysis, treatment of polysemy, increase of the number of extracted NEs, improvement on recall and standards compliance. The following subsections deal with these topics in detail.

<sup>5</sup>We do not consider synsets made up of instances as our aim is to extend synsets with instances and instances by definition cannot have instances.

<sup>6</sup>We use F-measure $_{\beta=0.5}$  instead of the most used F-measure $_{\beta=1}$  as we consider more important precision than recall for the current task.

### 3.1. Mapping Analysis

In the current research we use a database dump of the English Wikipedia from January 2008 and the LR WordNet 2.1. From this LR we consider the noun classes that contain instances. This leads us to a set of 893 synsets made up of 1012 monosemous words and 628 polysemous words. Table 1 shows the percentages of monosemous words, polysemous words and synsets that get mapped to Wikipedia categories by matching lemmas for three different dumps from April 2007, November 2007 and January 2008. As it can be seen, the continuous growth of Wikipedia allows us to increase the mapping percentage. 57.44% of the synsets were mapped to the April 2007 dump. This percentage increases to 60.02% for the November 2007 dump and to 60.58% for the January 2008 dump (the one we are currently working with).

Table 1: WordNet class nouns to Wikipedia categories mapping percentages

|                  |          | Wikipedia dump date |        |        |
|------------------|----------|---------------------|--------|--------|
|                  |          | 200704              | 200711 | 200801 |
| Monosem. Nouns   | Total    | 1012                |        |        |
|                  | Mapped   | 491                 | 509    | 518    |
|                  | Percent. | 48.51%              | 50.29% | 51.18% |
| Polysemous Nouns | Total    | 628                 |        |        |
|                  | Mapped   | 249                 | 265    | 262    |
|                  | Percent. | 39.64%              | 42.19% | 41.71% |
| Synsets          | Total    | 893                 |        |        |
|                  | Mapped   | 513                 | 536    | 541    |
|                  | Percent. | 57.44%              | 60.02% | 60.58% |

In order to get a better understanding on the mapping procedure, we have manually analysed a randomly selected set of WordNet classes which do not get mapped to any Wikipedia category. In most of the cases (75%), although there is not a matching category, there is a matching article in Wikipedia to which the class could be mapped. E.g. “oracle” could be mapped to the article “Oracle”. In 13% of the cases there is not a matching category and neither a matching article (e.g. “formal garden”). 10% of the times there is a matching category but the class is not mapped to it due to a PoS tagger error. E.g. the class “aquarium” is not mapped to the category “Aquaria” because the tagger fails to obtain “aquarium” as the lemma of the latter. The remaining 2% is due to having the class and matching category in different English variants. E.g. the class “railroad tunnel” (british) should be mapped to the category “railway tunnels” (american) but clearly their lemmas do not match.

### 3.2. Treatment of polysemy

Our first proposal only considered the mapping of monosemous words from WordNet. We hypothesise that instances could be useful to disambiguate WordNet polysemous words with respect to Wikipedia categories. For a polysemous word that is mapped to a category, we consider the instances for each of its senses and check if any of these instances is present also as an entry of the category. E.g. the word “obelisk” is mapped to the category “Obelisks”. It has two senses in WordNet (1. stone pillar, 2.

character used in printing). The first sense has one instance (“Washington Monument”) while the second has none. In the category “Obelisks” we find the instance “Washington Monument”. Thus, the sense chosen for the mapping would be the first one.

As the taxonomy of Wikipedia is usually deeper than that of WordNet, we consider not only looking for instances in the mapped categories but also in their hyponyms (subcategories). However, the subcategory relation in the categories taxonomy of Wikipedia does not always follow the hyponymy relation<sup>7</sup>. Therefore, in order to exploit subcategories, we need to check whether they are hyponyms or not. We propose to apply regular expression patterns which can hold both lexical and Part-of-Speech elements. If a subcategory matches a pattern then it is considered as a hyponym. From studying the category structure of Wikipedia, we come up with the following patterns (for each pattern we provide an example of matching subcategory for the category “Philosophers”):

- `^ category " by "`  
e.g. “philosophers by nationality”
- `^ category " of "`  
e.g. “philosophers of mind”
- `^ category " stubs" $`  
e.g. “philosophers stubs”
- `^ ((JJ|JJR|NN|NP)+ \ (CC (JJ|JJR|NN|NP)+) * \ " " category $`  
e.g. “Spanish philosophers”

As an example, we show how the word philosopher (1. specialist in philosophy, 2. wise person who is calm and rational) is disambiguated with respect to the category “Philosophers”. The first sense contains several instances like “Averroes” while the second contains none. “Averroes” is not present in the mapped category but it is in a subcategory that follows the hyponymy relation (“philosophers by nationality” → “Spanish philosophers”).

From a set of 262 polysemous words from WordNet which are lexically mapped to Wikipedia categories, this algorithm disambiguates 102 (39%). This low recall, which is due to the low number of instances present in WordNet, is compensated by a very high precision. In fact, all the disambiguated entries were correct. We analysed the reasons why 160 words (61%) were not disambiguated. There are two main cases:

- One of the senses from WordNet corresponds to the category but no common instance is found. This happens for 78% of the 160 words. For 74% of the words there is simply no common instance in both resources. For the remaining 4% a common instance does exist but it is in a subcategory that although being a hyponym of the mapped category, the algorithm is not

able to identify as such. E.g. “Colosseum, Amphitheatrum Flavium” is an instance of the second sense of “amphitheater”, which is mapped to the category “Amphitheaters”. “Colosseum” is present in the category “Roman amphitheatre buildings” which is a subcategory of “Amphitheaters”. However, the aforementioned patterns do not identify “Roman amphitheatre buildings” as a hyponym of “amphitheater”.

- No sense from WordNet corresponds to the category or the category has been changed. This occurs for the remaining 22% of the 160 words. An example of no sense corresponding to the mapped category happens for the word “assemblage” which has four senses: “a group of persons together in one place”, “a system of components assembled together for a particular purpose”, “the social act of assembling” and “several things grouped together or considered as a whole”. The mapped category, “Assemblage”, is “for assemblage artists”. As an example of a category change, the word “college” is mapped to the category “Colleges” but it has been moved to “Universities and colleges”. Obviously we cannot map “college” to “College and Universities” as by doing so we would end up with instances of universities under the class college.

### 3.3. Instance–class classification

Another modification to our initial approach concerns the instance–class classification. The aim is to increase recall (68.97%) without causing negative effects in precision (93.02%). We propose to take advantage of capitalisation norms as in the initial method but instead of looking for entry occurrences in the Web, we look for them in the body article of the entry, as in (Bunescu and Pasca, 2006). However, we do not only look for the entry in the article of the English Wikipedia, but in the article of several Wikipedias for other nine languages that follow these capitalisation norms (Catalan, Dutch, French, Italian, Norwegian, Portuguese, Romanian, Spanish and Swedish). This way the text size in which we look for occurrences of the entry is bigger and hence the results more representative. In order to obtain the entry title for each of these languages we use the multilingual links of Wikipedia that connect the same entry in different languages. This approach presents two advantages:

- Language independence. Whatever the language we are applying these procedures to, we can obtain the Wikipedia entry titles for languages which follow the aforementioned capitalisation norms.
- Avoidance of sense variation. A problem of the previous method is related to the fact that some nouns have senses in which they are instances and others in which they are classes. If an extracted entry is a NE but it has also a class sense the method could fail to classify it as a NE as in the Web we would find both senses. E.g. the Wikipedia entry “Children’s Machine” is a NE referring to a laptop developed by the OLPC (acronym of One Laptop Per Child). However, this term can also be found in the string “The children’s machine”, the title of book from Seymour Papert in which “children”

<sup>7</sup>E.g. In the category “Philosophers” there are subcategories that follow the hyponymy relation (e.g. “Philosophers by country”) but there are also others that do not (e.g. “Philosophy academics”).

and “machine” are classes. With the new method we look for “Children’s Machine” in the body of its article and so it is really unexpected to find this string referring to the book.

We have evaluated this new approach both by looking for entry occurrences only in the English Wikipedia and in the English Wikipedia plus the other nine aforementioned Wikipedias. Tables 2 and 3 present the results obtained respectively for each of the scenarios over a test set of 100 entries.

Table 2: Instance–class classification results using only the English Wikipedia

| Threshold | Precision | Recall | $F_{\beta=1}$ | $F_{\beta=0.5}$ |
|-----------|-----------|--------|---------------|-----------------|
| 0.81      | 73.24     | 89.66  | 80.62         | 78.00           |
| 0.83      | 72.86     | 87.93  | 79.69         | 77.27           |
| 0.85      | 72.86     | 87.93  | 79.69         | 77.27           |
| 0.87      | 73.91     | 87.93  | 80.32         | 78.06           |
| 0.89      | 73.91     | 87.93  | 80.32         | 78.06           |
| 0.91      | 73.91     | 87.93  | 80.32         | 78.06           |
| 0.93      | 73.91     | 87.93  | 80.32         | 78.06           |
| 0.95      | 73.91     | 87.93  | 80.32         | 78.06           |

Table 3: Instance–class classification results using Wikipedia for ten languages

| Threshold | Precision | Recall | $F_{\beta=1}$ | $F_{\beta=0.5}$ |
|-----------|-----------|--------|---------------|-----------------|
| 0.81      | 77.62     | 89.66  | 83.2          | 81.26           |
| 0.83      | 77.62     | 89.66  | 83.2          | 81.26           |
| 0.85      | 78.79     | 89.66  | 83.88         | 82.11           |
| 0.87      | 78.79     | 89.66  | 83.88         | 82.11           |
| 0.89      | 79.69     | 87.93  | 83.61         | 82.26           |
| 0.91      | 79.69     | 87.93  | 83.61         | 82.26           |
| 0.93      | 79.37     | 86.21  | 82.64         | 81.53           |
| 0.95      | 79.37     | 86.21  | 82.64         | 81.53           |

The best F-measure $_{\beta=0.5}$  is obtained for the thresholds 0.87 to 0.95 when using only the English Wikipedia (78.06%) and for the thresholds 0.89 to 0.91 when using ten Wikipedias (82.26%). Therefore the best threshold value could be set to 0.91, which is exactly the same value for our previous approach. For this threshold, using more text allow us to obtain better precision (79.69% vs. 73.91%) and the same recall (87.93%), which supports our hypothesis of using different Wikipedias to increase the text size. Compared to our previous approach (web search) the current approach obtains 21.56% higher recall (87.93% vs. 68.97%) but 14.33% lower precision (79.69% vs. 93.02%). By analysing the results, we have found a drawback of the current approach compared to web search. The number of occurrences found per article is quite low: 6.03 when using only the English Wikipedia and 9.77 when using also the others. These values contrast with those obtained for the web search. In fact, for that experiment we set the number of occurrences per article to 100 and found such a high number for all the 100 articles of the test set.

We conclude then that the advantages of both approaches could be combined by extracting significant terms from the entry body text in Wikipedia (e.g. by applying the tf-idf measure) and search in the Web pages in which the entry title and these terms appear. Following with the example of the entry “Children’s Machine”, from the ten first results from Google six correspond to the computer and the remaining four to Papert’s book. However, if we extract significant terms from the body text of the Wikipedia entry such as “OLPC” and “\$100 laptop”, and we search in Google the three terms, then all the first 10 results correspond to the computer.

### 3.4. Increase of the number of extracted NEs

Regarding the number of NEs added to WordNet, it can be easily boosted just by extracting not only the NE entries that belong to the mapped categories but also the NE entries that belong to the subcategories that fulfil the hyponymy relation (see subsection 3.2.).

We have extracted NEs from Wikipedia for the 541 mapped synsets (see table 1) both considering hyponym subcategories and not considering them. Table 4 provides quantitative data about the NEs extracted. For each case, we show not only the number of NEs which is added to WordNet but also the amount of orthographic variants (written forms) of these NEs and the number of instance relations that WordNet is enriched with. By considering subcategories we are able to enrich WordNet with more than 300,000 NEs and more than 380,000 instance relations.

Table 4: Extracted NEs

|                    | Without subcategories | With subcategories |
|--------------------|-----------------------|--------------------|
| NEs                | 16,328                | 310,742            |
| Written forms      | 16,672                | 452,017            |
| Instance relations | 26,438                | 381,043            |

Table 5 provides results about the nature of the NEs added to WordNet. It shows the number of instances added to the different noun lexicographic files of WordNet. For each lexicographic file to which instances are added we include an example of such an instance together with the synset it is attached to.

### 3.5. Standard-compliant Output

Finally, in order to make the procedures independent from specific LRs we provide an output format compliant to standards. The elements that are part of this output are mainly NEs, orthographic variants of these NEs and classes to which these NEs belong (by means of “instance of” relations). Due to the fact that this data could be naturally represented by means of a LR and because the final aim is to extend a LR with this information we have decided to follow the Lexicon Markup Framework (LMF), an ISO standard for the representation of lexicons (Francopoulo et al., 2008) (ISO 24613, 2008), in order to encode the output.

We have developed a NE repository as a database whose structure is compliant with LMF. The idea is to insert the

Table 5: Number of NEs per lexicographic file

| Lexicographic File | Number of NEs | Example   |
|--------------------|---------------|---|
| act                | 4,214         | Project_Pluto instanceOf project0_4                     |
| animal             | 1             | Power_Animal_(Gaoranger) instanceOf fictional_animal0_5 |
| artifact           | 23,878        | Akinada_Bridge instanceOf suspension_bridge0_6          |
| communication      | 1,973         | Flower_of_Scotland instanceOf national_anthem0_10       |
| event              | 58            | Sino-Soviet_split instanceOf schism0_11                 |
| group              | 1,216         | Medici instanceOf family0_14                            |
| location           | 43,582        | Incense_Route instanceOf trade_route0_15                |
| object             | 28,180        | Pyxis instanceOf constellation0_17                      |
| person             | 277,941       | Vladimir_Kotelnikov instanceOf electrical_engineer0_18  |

extracted NEs into this repository and to have modules for each specific LR in order to convert the information from the LMF compliant database to the specific format of the LR. For the current research in which the target LR to be enriched is WordNet, we have developed a module which converts the information regarding the NEs of the repository to the specific format of the lexicographic files of WordNet (grind).

As an example, we include both the XML LMF code (see Figure 1), the database entries (see Figure 2) and the WordNet specific format (see Figure 3) for the information extracted regarding the NE “Tim Robbins”.

#### 4. Conclusions and Future Work

This paper has presented Named Entity WordNet, a NE rich LR for English. Several aspects regarding the construction of this resource have been discussed in detail. These include the treatment of polysemous nouns, an analysis of mapped classes and of the results and the design of a standard compliant output format. The resulting resource contains 310,742 NEs and 381,043 “instance of” relations and is publicly available<sup>8</sup>.

It is important to mention also that NEWN has been automatically built by applying a generic methodology. In fact, our approach could be applied to any LR that has a noun taxonomy. The only adjustment that needs to be done when applying this method to a different LR is to develop a module that converts the generic standard compliant format of the NE repository into the format of the specific LR.

<sup>8</sup>NEWN can be downloaded at <http://dlsi.ua.es/~atoral/#Resources>

Moreover, the method has a high degree of language independence; the only language dependent aspect is the set of lexical and Part-of-Speech patterns used for identifying hyponym subcategories.

Finally we sketch some future work lines. First, we will apply the presented methodology to another LR for another language to empirically prove the generic nature of our approach. The other two lines regard increasing mapping recall. On one hand, as in most of the cases in which a class is not mapped to any category it could be mapped against a Wikipedia entry (74%), we plan to exploit article content from Wikipedia, e.g. by applying state-of-the-art Information Extraction techniques in order to extract instances from the article body. On the other, we plan to apply Textual Entailment techniques between classes’ glosses and articles’ abstracts in order to increase the disambiguation recall and therefore the mapping recall for polysemous classes.

#### 5. References

- Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, Italy*, pages 9–16, April.
- Claude de Loupy, Eric Crestan, and Elise Lemaire. 2004. Proper Nouns Thesaurus for Document Retrieval and Question Answering. In *Atelier Question-Réponse, Traitement Automatique des Langues Naturelles (TALN)*.
- G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. 2008. (forthcoming) Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation Journal*.
- ISO 24613. 2008. Languages Resources Management – Lexical Markup Framework (LMF), rev.15 ISOTC37SC4 FDIS. [Online; accessed 25-March-2008].
- Cvetana Krstev, Důsko Vitas, Denis Maurel, and Mickaël Tran. 2005. Multilingual Ontology of Proper Names. In *Proceedings of the Language and Technology Conference*, pages 116–119.
- D. Levy, R. Catizone, B. Battacharia, A.Krotov, and Y. Wilks. 1997. CONVERSE: A conversational companion. In *Proceedings of the 1st International Workshop on Human-Computer Conversation*, Bellagio, Italy.
- G. Mann. 2002. Fine-grained proper noun ontologies for question answering.
- G. A. Miller and F. Hristea. 2006. Wordnet nouns: Classes and instances. *Computational Linguistics*, 32(1):1–3.
- G. A. Miller. 1995. Wordnet: A lexical database for english. *Communications of ACM*, (11):39–41.
- Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended Named Entity Hierarchy. In *Proceedings of Third International Conference on Language Resources and Evaluation*.
- Svetlana Sheremetyeva, Jim Cowie, Sergei Nirenburg, and Rémi Zajac. 1998. Multilingual Onomasticon as a Multipurpose NLP Resource. In *Proceedings of the First International Conference on Language Resources and Evaluation*.

Figure 1: XML LMF code for the NE “Tim Robbins”

```

<Lexicon>
  <feat att="name" val="NamedEntityRepository"/>
  <feat att="language" val="en"/>
  <LexicalEntry id="LE_Tim_Robbins">
    <feat att="POS" val="ProperNoun"/>
    <Lemma>
      <FormRepresentation>
        <feat att="writtenform" val="Tim Robbins"/>
        <feat att="VariantType" val="full"/>
      </FormRepresentation>
      <FormRepresentation>
        <feat att="writtenform" val="Timothy_Francis_Robbins"/>
        <feat att="VariantType" val="alias"/>
      </FormRepresentation>
    </Lemma>
    <Sense id="S_Tim_Robbins">
      <feat att="Resource" val="Wikipedia"/>
      <feat att="ResourceId" val="269416"/>
      <SemanticDefinition>
        <feat att="text" val="[...] is an American Academy Award-winning [...]"/>
      </SemanticDefinition>
      <SenseRelation targets="S_screenwriter">
        <feat att="semanticrelation" val="instance_of"/>
      </SenseRelation>
      <SenseRelation targets="S_film_director">
        <feat att="semanticrelation" val="instance_of"/>
      </SenseRelation>
    </Sense>
  </LexicalEntry>
  <LexicalEntry id="LE_screenwriter">
    <feat att="POS" val="Noun"/>
    <Lemma id="LM_screenwriter"> [...] </Lemma>
    <Sense id="S_screenwriter0_18">
      <feat att="Resource" val="WordNet"/>
      <feat att="ResourceId" SourceVal="noun.person:screenwriter0"/>
    </Sense>
  </LexicalEntry>
  <LexicalEntry ID="LE_film_director">
    <feat att="POS" val="Noun"/>
    <Lemma id="LM_film_director"> [...] </Lemma>
    <Sense id="S_film_director0_18">
      <feat att="Resource" val="WordNet"/>
      <feat att="ResourceId" SourceVal="noun.person:film_director0"/>
    </Sense>
  </LexicalEntry>
</Lexicon>

```

Beth M. Sundheim, Scott Mardis, and John Burger. 2006. Gazetteer Linkage to WordNet. In *Proceedings of the Third International WordNet Conference*, pages 103–104.

Antonio Toral and Rafael Muñoz. 2007. Towards a Named Entity Wordnet (NEWN). In *Proceedings of the 6th International Conference on Recent Advances in Natural Language Processing*, pages 604–608, September.

Mickaël Tran, Thierry Grass, and Denis Maurel. 2004. An

Ontology for Multilingual Treatment of Proper Names. In *Proceedings of OntoLex 2004*.

Figure 2: Database entries for the NE “Tim Robbins”

Table LexicalEntry

| le_id         | le_pos |
|---------------|--------|
| film_director | N      |
| screenwriter  | N      |
| Tim_Robbins   | PN     |

Table FormRepresentation

| le_id       | fr_written_form         | fr_variant_type |
|-------------|-------------------------|-----------------|
| Tim_Robbins | Timothy_Francis_Robbins | alias           |
| Tim_Robbins | Tim_Robbins             | full            |

Table Sense

| s_id              | le_id         | resource  | resource_id                    | definition  |
|-------------------|---------------|-----------|--------------------------------|-------------|
| film_director0_18 | film_director | WordNet   | noun.person:<br>film_director0 |             |
| screenwriter0_18  | screenwriter  | WordNet   | noun.person:<br>screenwriter0  |             |
| Tim_Robbins       | Tim_Robbins   | Wikipedia | 269416                         | .. is an .. |

Table SenseRelation

| s_source_id | s_target_id       | sr_relation |
|-------------|-------------------|-------------|
| Tim_Robbins | film_director0_18 | instanceOf  |
| Tim_Robbins | screenwriter0_18  | instanceOf  |

Figure 3: WordNet specific code for the NE “Tim Robbins”

```
{ Tim_Robbins, Timothy_Francis_Robbins,
  noun.person:screenwriter,@i noun.person:film_director,@i
  ([...] is an American Academy Award-winning [...]) }
```