Wikipedia as Frame Information Repository

Sara Tonelli
FBK-irst
I-38100, Trento, Italy
satonelli@fbk.eu

Claudio Giuliano
FBK-irst
I-38100, Trento, Italy
giuliano@fbk.eu

Abstract

In this paper, we address the issue of automatic extending lexical resources by exploiting existing knowledge repositories. In particular, we deal with the new task of linking FrameNet and Wikipedia using a word sense disambiguation system that, for a given pair frame – lexical unit (F, l), finds the Wikipage that best expresses the the meaning of l. The mapping can be exploited to straightforwardly acquire new example sentences and new lexical units, both for English and for all languages available in Wikipedia. In this way, it is possible to easily acquire good-quality data as a starting point for the creation of FrameNet in new languages. The evaluation reported both for the monolingual and the multilingual expansion of FrameNet shows that the approach is promising.

1 Introduction

Many applications in the context of natural language processing or information retrieval have proved to convey significant improvement by exploiting lexical databases with high-quality annotation such as *FrameNet* (Fillmore et al., 2003) and *WordNet* (Fellbaum, 1998). Nevertheless, the practical use of similar resources is often biased by their limited coverage because manual annotation is time-consuming and requires a relevant financial effort. For this reason, some research activities have focused on the automatic enrichment of such resources with annotated information in (near) manual quality. The main strategy proposed was the mapping between resources in order to reciprocally enrich different lexical databases by linking their information layers. This has proved to be useful in several tasks, from verb classification (Chow and Webster, 2007) to semantic role labeling (Giuglea and Moschitti, 2006), open text semantic parsing (Shi and Mihalcea, 2004) and textual entailment (Burchardt and Frank, 2006).

In this work, we focus on the automatic enrichment of the FrameNet database for English and we propose a new framework to extend this procedure to new languages. While similar works in the past have mainly proposed to automatically extend the FrameNet database by mapping frames and WordNet synsets (Shi and Mihalcea (2005), Johansson and Nugues (2007), and Tonelli and Pighin (2009)), we present an explorative approach that for the first time exploits Wikipedia to this purpose. In particular, given a lexical unit l belonging to a frame F, we devise a strategy to link l to the Wikipedia article that best captures the sense of l in F. This is basically a word disambiguation (WSD) problem (Erk, 2004) and to this purpose we employ a state-of-the-art WSD system (Gliozzo et al., 2005). The mapping between (F, l) pairs and Wikipedia pages could then be exploited for three further subtasks: (a) automatically extract from Wikipedia all sentences pointing to the Wikipage mapped with (F, l) and assign them to F; (b) automatically expand the lexical units sets in the English FrameNet by exploiting the redirecting and linking strategy of Wikipedia; and (c) since Wikipedia is available in 260 languages, use the English Wikipedia article linked to (F, l) as a bridge to carry out sentence and lexical unit retrieval in other languages. The set of automatically collected data would represent the starting point for the creation of FrameNet in new languages. In fact, having a repository of sentences extracted from Wikipedia which have already been divided by sense would significantly speed up the annotation process. In this way, the annotators would not need to extract all sentences in a corpus containing l and classify them by sense. Instead, they should simply validate the given sentences and assign the correct frame elements.

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 276–285,
Singapore, 6-7 August 2009.

In the following, we start by providing a brief overview of FrameNet and Wikipedia and we present their structure and organization. Next, we describe the algorithm for mapping lexical units and Wikipages and the word sense disambiguation algorithm employed by the system. In Section 5 we describe the dataset used in the first experiment and report evaluation results of the mapping between (F, l) pairs and Wikipedia senses. In Section 6 we describe an application of the mapping, i.e. the automatic enrichment of English FrameNet. We describe the data extraction process and evaluate the quality of the data. In Section 7 we describe and evaluate another application of the mapping, i.e. the acquisition of data for the automatic creation of Italian FrameNet using the Italian Wikipedia. Finally, we draw conclusions and present future research directions.

2 FrameNet and Wikipedia

FrameNet (Fillmore et al., 2003) is a lexical resource for English based on corpus evidence, whose conceptual model comprises a set of prototypical situations called *frames*, the frame-evoking words or expressions called *lexical units* (LUs) and the roles or participants involved in these situations, called *frame elements*. All lexical units belonging to the same frame have similar semantics but, differently from WordNet synsets, they can belong to different categories and present different parts of speech. For example, the KILLING frame is described in the FrameNet database[1] as *"A Killer or Cause causes the death of the Victim"*. The elements in capitals are the semantic roles (*frame elements*) typically involved in the KILLING situation. The frame definition comes also with the list of frame-evoking lexical units, namely *annihilate.v, annihilation.n, butchery.n, carnage.n, crucify.v, deadly.a*, etc. Since FrameNet is a corpus-based resource, every lexical unit should be instantiated by a set of example sentences, where the frame elements are annotated as well. Instead, FrameNet is still an ongoing project and in the latest release (v. 1.3) there are about 3,380 lexical units out of 10,195 that come with no example sentences. In this work we focus on these lexical units and propose how to automatically collect the missing sentences. Anyhow, the algorithm we propose is suitable also for expanding sentence sets already present in FrameNet.

Wikipedia[2] is one of the largest online repositories of encyclopedic knowledge, with millions of articles available for a large number of languages (>2,800,000 for English). The *article* (or *page*) is the basic entry in Wikipedia. Every article has an unique reference, i.e., one or more words that identify the page and are present in its URL. For example, *Ball_(dance)* identifies the page that describes several types of ball intended as formal dance, while *Dance_(musical_form)* describes the dance as musical genre. Every Wikipedia article is linked to others, and in the body of every page there are plenty of links to connect the most relevant terms to other pages. Another important attribute is the presence of about 3,000,000 redirection pages, that given an identifier that is not present in Wikipedia, automatically display the page with the most semantically similar identifier (for example *Killing* is redirected to the *Murder* page). Wikipedia contains also more than 100,000 disambiguation pages listing all senses (pages) for an ambiguous entity. For example, *Book* has 9 senses, which correspond to 9 different articles. Wikipedia structure and quality make this resource particularly suitable for information extraction and word sense disambiguation tasks (Csomai and Mihalcea (2008) and Milne and Witten (2008)). In fact, page references can be seen as senses and Wikipedia as a large sense inventory. From this point of view, also linking a lexical unit to the correct Wikipedia page is a word sense disambiguation issue because it implies recognizing what meaning the lexical unit has in the given frame. For example, *dance.n* in the SOCIAL_EVENT frame should be linked to *Ball_(dance)* and not to *Dance_(musical_form)*.

3 The Mapping Algorithm

In this section, we describe how to map a frame – lexical unit pair (F, l) into the Wikipedia article that best captures the sense of l as defined in F. The mapping problem is casted as a supervised WSD problem, in which l must be disambiguated using F to provide the context and Wikipedia to provide the sense inventory and the training data. Even if the idea of using Wikipedia links for disambiguation is not novel (Cucerzan, 2007), it is applied for the first time to FrameNet lexical units, considering a frame as a sense definition. The proposed algorithm is summarized as follows:

[1] http://framenet.icsi.berkeley.edu

[2] http://en.wikipedia.org

Step 1 For each lexical unit l, we collect from the English Wikipedia dump[3] all contexts[4] where l is the anchor of an internal link (wiki link). The set of targets represents the senses of l in Wikipedia and the contexts are used as labelled training examples. For example, the lexical unit *building.n* in the frame *Buildings* is an anchor in 708 different contexts that point to 42 different Wikipedia pages (senses).

Step 2 The set of contexts with their corresponding senses is then used to train the WSD system described in Section 4. For example, the context "The *building*, which date from the mid-to-late 19th century, were built in a variety of High Victorian architectural styles." is a training example for the sense defined by the Wikipedia page *Building*.

Step 3 Finally, the disambiguation model learned in the previous step is used to map a pair (F, l) to a Wikipedia article. (F, l) is represented as a fictitious-context derived by aggregating the frame definition and all lexical units associated to F. We used the term "fictitious-context" to remark the slight difference in structure compared with the training contexts (i.e., the Wikipedia paragraphs). For example, "... structures forming an enclosure and providing protection from the elements ... acropolis arena auditorium bar building ..." is the fictitious-context built for the pair *(Buildings, building.n)*. The sense, i.e., the Wikipedia article, assigned to the fictitious-context by the disambiguation algorithm uniquely defines the mapping. The previous example is assigned to the Wikipedia page *Building*.

4 The WSD Algorithm

Gliozzo et al. (2005) proposed an elegant approach to WSD based on kernel methods. The algorithm proved effective at Senseval-3 (Mihalcea and Edmonds, 2004) and, nowadays, it still represents the state-of-the-art in WSD (Pradhan et al., 2007). Specifically, they addressed these issues: (i) independently modeling domain and syntagmatic aspects of sense distinction to improve feature representativeness; and (ii) exploiting external knowledge acquired from unlabeled data, with the purpose of drastically reducing the amount of labeled training data. The first direction is based on the linguistic assumption that syntagmatic and domain (associative) relations are crucial for representing sense distictions, but they are originated by different phenomena. Regarding the second direction, it is possible to obtain a more accurate prediction by taking into account unlabeled data relevant for the learning problem (Chapelle et al., 2006).

On the other hand, kernel methods are theoretically well founded in statistical learning theory and shown good empirical results in many applications (Shawe-Taylor and Cristianini, 2004). The strategy adopted by kernel methods consists of splitting the learning problem into two parts. They first embed the input data in a suitable feature space, and then use a linear algorithm (e.g., support vector machines) to discover nonlinear patterns in the input space. The kernel function is the only task-specific component of the learning algorithm. Thus, to develop a WSD system, one only needs to define appropriate kernel functions to represent the domain and syntagmatic aspects of sense distinction and to exploit the properties of kernel functions in order to define a composite kernel that combines and extends individual kernels.

The WSD system described in the following consists of a composite kernel (Section 4.3) that combines the domain and syntagmatic kernels. The former (Section 4.1) models the domain aspects of sense distinction, the latter (Section 4.2) represents the syntagmatic aspects of sense distinction.

4.1 Domain Kernel

It is been shown that domain information is fundamental for WSD (Magnini et al., 2002). For instance, the (domain) polysemy between the computer science and the medicine senses of the word "virus" can be solved by considering the domain of the context in which it appears.

In the context of kernel methods, domain information can be exploited by defining a kernel function that estimates the domain similarity between the contexts of the word to be disambiguated. The simplest method to estimate the domain similarity between two texts is to compute the cosine similarity of their vector representations in the vector space model (VSM). The VSM is a k-dimensional space $\mathbb{R}^k$, in which the text t_j is represented by a vector $\vec{t_j}$, where the i^{th} component is the term

[3] http://download.wikimedia.org/enwiki/20090306

[4] A context corresponds to a line of text in the Wikipedia dump and it is represented as a paragraph in a Wikipedia article.

frequency of the term w_i in t_j. However, such an approach does not deal well with lexical variability and ambiguity. For instance, despite the fact that the sentences "he is affected by AIDS" and "HIV is a virus" express concepts closely related, their similarity is zero in the VSM because they have no words in common (they are represented by orthogonal vectors). On the other hand, due to the ambiguity of the word "virus" , the similarity between the sentences "the laptop has been infected by a virus" and "HIV is a virus" is greater than zero, even though they convey very different messages.

To overcome this problem, Gliozzo et al. (2005) introduced the domain model (DM) and show how to define a domain VSM in which texts and terms are represented in a uniform way. A DM is composed of soft clusters of terms. Each cluster represents a semantic domain, that is, a set of terms that often co-occur in texts having similar topics. A DM is represented by a $k \times k'$ rectangular matrix $\mathbf{D}$, containing the degree of association among terms and domains.

The matrix $\mathbf{D}$ is used to define a function $\mathcal{D} : \mathbb{R}^k \rightarrow \mathbb{R}^{k'}$, that maps the vector $\vec{t_j}$ represented in the standard VSM, into the vector $\vec{t'_j}$ in the domain VSM. $\mathcal{D}$ is defined by

$$\mathcal{D}(\vec{t_j}) = \vec{t_j}(\mathbf{I^{IDF}D}) = \vec{t'_j}, \tag{1}$$

where $\vec{t_j}$ is represented as a row vector, $\mathbf{I^{IDF}}$ is a $k \times k$ diagonal matrix such that $i^{IDF}_{i,i} = IDF(w_i)$, and $IDF(w_i)$ is the inverse document frequency of w_i.

In the domain space, the similarity is estimated by taking into account second order relations among terms. For example, the similarity of the two sentences "He is affected by AIDS" and "HIV is a virus" is very high, because the terms AIDS, HIV and virus are strongly associated with the domain medicine.

Singular valued decomposition (SVD) is used to acquire in a unsupervised way the DM from a corpus represented by its term-by-document matrix $\mathbf{T}$. SVD decomposes the term-by-document matrix $\mathbf{T}$ into three matrixes $\mathbf{T} \simeq \mathbf{V}\mathbf{\Sigma_{k'}}\mathbf{U}^T$, where $\mathbf{V}$ and $\mathbf{U}$ are orthogonal matrices (i.e., $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I}$) whose columns are the eigenvectors of $\mathbf{T}\mathbf{T}^T$ and $\mathbf{T}^T\mathbf{T}$ respectively, and $\mathbf{\Sigma_{k'}}$ is the diagonal $k \times k$ matrix containing the highest $k' \ll k$ eigenvalues of $\mathbf{T}$, and all the remaining elements set to 0. The parameter k' is the dimensionality of the domain VSM and can be fixed in advance. Under this setting, the domain matrix $\mathbf{D}$ is defined by

$$\mathbf{D} = \mathbf{I^N}\mathbf{V}\sqrt{\mathbf{\Sigma_{k'}}} \tag{2}$$

where $\mathbf{I^N}$ is a diagonal matrix such that $\mathbf{i^N_{i,i}} = \frac{1}{\sqrt{\langle \vec{w'_i}, \vec{w'_i} \rangle}}$, $\vec{w'_i}$ is the i^{th} row of the matrix $\mathbf{V}\sqrt{\mathbf{\Sigma_{k'}}}$. The domain kernel is explicitly defined by

$$K_D(t_i, t_j) = \langle \mathcal{D}(t_i), \mathcal{D}(t_j) \rangle, \tag{3}$$

where $\mathcal{D}$ is the domain mapping defined in Equation 1. Finally, the domain kernel is further extended to include the standard bag-of-word kernel.

4.2 Syntagmatic Kernel

Kernel functions are not restricted to operate on vectorial objects $\vec{x} \in \mathbb{R}^k$. In principle, kernels can be defined for any kind of object representation, such as strings and trees. As syntagmatic relations hold among words collocated in a particular temporal order, they can be modeled by analyzing sequences of words. Therefore, the string kernel (Shawe-Taylor and Cristianini, 2004) is a valid tool to represent such relations. It counts how many times a (non-contiguous) subsequence of symbols u of length n occurs in the input string s, and penalizes non-contiguous occurrences according to the number of gaps they contain. Formally, let V be the vocabulary, the feature space associated with the string kernel of length n is indexed by a set I of subsequences over V of length n. The (explicit) mapping function is defined by

$$\phi^n_u(s) = \sum_{\mathbf{i}:u=s(\mathbf{i})} \lambda^{l(\mathbf{i})}, u \in V^n, \tag{4}$$

where $u = s(\mathbf{i})$ is a subsequence of s in the positions given by the tuple $\mathbf{i}$, $l(\mathbf{i})$ is the length spanned by u, and $\lambda \in]0, 1]$ is the decay factor used to penalize non-contiguous subsequences.

The associated string kernel is defined by

$$K_n(s_i, s_j) = \langle \phi^n(s_i), \phi^n(s_j) \rangle = \sum_{u \in V^n} \phi^n(s_i)\phi^n(s_j) \tag{5}$$

Gliozzo et al. (2005) modified the generic definition of the string kernel in order to take into account (sparse) collocations. Specifically, they defined syntagmatic kernels as a combination of string kernels applied to sequences of words in a fixed-size window centered on the word to be disambiguated. This formulation allows estimating the number of common (sparse) subsequences of

words (i.e., collocations) between two examples, in order to capture syntagmatic similarity. The syntagmatic kernel is defined by

$$K_S(s_i, s_j) = \sum_{n=1}^{p} K_n(s_i, s_j), \quad (6)$$

where K_n is the string kernel defined in Equation 5 and the parameter n represents the length of the subsequences analyzed when estimating the similarity between contexts. Notice that the syntagmatic kernel is only effective for those fictitious contexts in which the lexical units do occur in meaningful sentences, however this is not guaranteed for the lexical units without examples.

4.3 Composite Kernel

Finally, to combine domain and syntagmatic information, the composite kernel is defined by

$$K_{WSD}(t_i, t_j) = \hat{K}_D(t_i, t_j) + \hat{K}_S(t_i, t_j), \quad (7)$$

where $\hat{K}_D$ and $\hat{K}_S$ are normalized kernels defined in Equation 3 and 6, respectively.[5] It follows directly from the explicit construction of the feature space and from closure properties of kernels that it is a valid kernel.

5 Mapping task

In this section we report the first experiment, namely the mapping between (F, l) pairs and a Wikipedia pages. We describe the experimental setup and then present the corresponding evaluation.

5.1 Experimental setup

We applied our algorithm to all lexical units that do not have any example sentence in the FrameNet database. In principle, the proposed approach can be applied to every lexical unit, and we expect the algorithm performance to improve if some example sentences are already available because they could be added to the fictitious-context used to represent (F, l) in the system. Nevertheless, in this explorative study we wanted to focus on the harder cases, even if results are likely to be worse than on the whole FrameNet database.

In FrameNet, 3,305 (F, l) pairs have no example sentences (536 pairs with adjectival LU, 1313 verbal LU, 1456 nominal LU). Since Wikipedia is basically a resource organized by concepts, which are generally expressed by nouns, we decided to restrict our experiment to nominal lexical units. Besides, many verbal and adjectival concepts in Wikipedia are redirected to nominal identifiers. So, we randomly selected 900 pairs with nominal lexical units. For the moment, we decided to discard lexical units expressed by multiwords (about 150), which will be taken into account in a future version of our system. The average ambiguity of the 900 LUs considered is 1.24 in FrameNet. Instead, every LU corresponds to about 35 candidate senses in Wikipedia.

In order to perform WSD, we built the domain model from the 200,000 most visited Wikipedia articles. After removing terms that occur less than 5 times, the resulting dictionaries contain about 300,000 terms. We used the SVDLIBC package[6] to compute the SVD, truncated to 100 dimensions. The experiments were performed using the SVM package *LIBSVM* (Chang and Lin, 2001) customized to embed the kernels described in Section 4.

5.2 Evaluation

In this first evaluation step, we focus on the quality of the mapping between (F, l) pairs and Wikipedia articles. In order to evaluate the system output, we created a gold standard where 250 (F, l) pairs randomly extracted from the nominal subset described above have been manually linked to the Wikipedia page (if available) that best corresponds to the meaning of l in F. The pairs have been chosen in order to maximize the frame variability, i.e. every pair corresponds to a different frame. Since our gold standard contains 34% of all frames in the FrameNet database, we believe that, despite its limited size, it is well representative of FrameNet characteristics. Evaluation was carried out comparing the system output against the gold standard. Results are reported in Table 1. The baseline was computed considering the *most frequent sense* of every lexical unit in Wikipedia. This element is obtained by taking into account all occurrences in Wikipedia where the lexical unit LU we consider is anchored to a given page. The most frequent sense for LU is the page to which LU is most frequently linked in Wikipedia. Since about 14% of the lexical units in the gold standard are not present in Wikipedia, we also estimated an upper bound accuracy of 0.86. This confirms our in-

[5] $\hat{K}(x_i, x_j) = \frac{K(x_i, x_j)}{\sqrt{K(x_j, x_j)K(x_i, x_i)}}$

[6] http://tedlab.mit.edu/~dr/svdlibc/

tuition that FrameNet and Wikipedia are linkable resources to a large extent and that our task is well-founded.

	Accuracy
Baseline	0.66
System output	**0.71**
Upper bound	0.86

Table 1: Accuracy evaluation.

Wrong assignments include also problematic cases that are not directly connected to proper system errors. One of the most relevant issues is the different granularity between FrameNet frames and Wikipages. For example, the NETWORK frame is defined as *"a set of entities of the same or similar types (Nodes) are linked to each other by Connections to form a Network allowing for the flow of information, resources, etc."*. Even if the listed lexical units (*network.n* and *web.n*) and some examples refer to the informatics domain, the situation described in the FrameNet database is more general. Wikipedia instead lists several pages that may be seen as subdomains of NETWORK such as *Computer_network*, *Social_network*, *Telecommunications_network*, etc. In the future, it may be worth modifying the system in order to allow multiple assignments of Wikipages for every frame.

In other cases, frame definitions seem not to be very consistent and it is very difficult to discriminate between two frames even for a human annotator. For example, ESTIMATED_VALUE and ESTIMATING include both *estimation.n* as lexical unit, but since their frame definitions are almost the same and the other lexical units in the same frame are not discriminative, the system links both (F, l) pairs to the same Wikipedia article.

6 English FrameNet expansion

In the following part of the experiment, we want to investigate to what extent the FrameNet – Wikipedia mapping can be effectively applied to automatically expand the FrameNet database with new example sentences, and eventually to acquire new lexical units. For every (F, l) pair, we consider the linked Wikipedia sense *s* and extract all sentences C_s in Wikipedia with a reference to s. In this way, we can assume that, if s was linked to (F, l), C_s can be included in the example sentences of F. This repository of sentences is already divided by sense and can significantly speed-up manual annotation. On the other hand, the extracted sentences could enrich the training set of machine learning systems for frame annotation to improve the frame identification step. In fact, this task has raised growing interest in the NLP community, with a devoted challenge at the last SemEval campaign (Baker et al., 2007).

This retrieval process allows also to extract from C_s all words W_s that have an embedded reference to s in the form *word*. In this way, W_s are automatically included in F as new lexical units. In this phase, redirecting links are very useful because they automatically connect a word or expression to its nearest sense in case there is no specific page for this word. The information about redirecting allows also to account for orthographic variations of the same lexical unit, for example *collectible* is redirected to *collectable*.

We explain the data extraction process in the light of an example from our dataset. Our WSD system assigned to the (F, l) pair (WORD_RELATIONS – *homonym.n*) the Wikipage `http://en.wikipedia.org/wiki/Homonym`. So, we extracted from the English Wikipedia dump all sentences where the anchor *<a href="/wiki/Homonym"...* > appears and assumed that the word or multiword expression that is linked to the *Homonym* site may be a good candidate as lexical unit for the WORD_RELATIONS frame. In this case, the example sentences were 186. Apart from *homonym*, the candidate lexical units are *homograph*, *homophone*, *homophonous*, *homonymic*, *heteronym*, *same*. Among them, only the latter is not appropriate, even if the sentence where it occurs is semantically connected to the WORD_RELATIONS frame: "In Hebrew the word 'thus' has the *same* triconsonantal root". Instead, *homonymic* and *heteronym* can be acquired as new lexical units for WORD_RELATIONS, and *homograph*, for which no example sentences are provided in FrameNet, can be automatically instantiated by a set of examples.

6.1 Experimental setup

We considered 893 frame – lexical unit pairs assigned to Wikipedia pages following the algorithm described in Section 3. We discarded 7 pairs for which the system reported an assignment failure, i.e. the best sense delivered is the disambigua-

tion page. Then we extracted a set of sentences for every (F, l) pair as described in the previous paragraph. Statistics about the retrieved data is reported in Table 2.

	English Wikipedia
(F, l) pairs	893
N. of extracted sents	964,268
Avg. sents per (F, l)	1,080

Table 2: Extracted data from English Wikipedia

6.2 Evaluation

The dimension of the extracted corpus does not allow to carry out a comprehensive evaluation. For this reason, we manually evaluated 1,000 sentences, i.e. we considered 20 (F, l) pairs, and for each of them we evaluated 50 sentences extracted from our large repository. Both (F, l) pairs and the assigned sentences were randomly selected. In particular, the 20 (F, l) pairs do not contain only correctly assigned pairs, in fact three of them are wrong. Anyhow, the 20 pairs seem to be a representative subset of the 893 pairs considered in our experiment because they include both monosemic lexical units (*gynaecology.n* in MEDICAL_SPECIALTIES) and more ambiguous ones (*club.n* in the WEAPON frame).

Our evaluation shows that 78% of the sentences were correctly linked to (F, l) pairs. This value is higher that the mapping accuracy between (F, l) and Wikipages reported in Section 5.2. In fact, we noticed that even if the Wikipage assigned to (F, l) is not the article that best corresponds to the meaning of l in F, some sentences pointing to it may be appropriate to express l.

As we already mentioned in Section 5.2, the different granularity of the information encoded by frames and Wikipages impacts on the output quality. For example, *conversion.n* in CAUSE_CHANGE has a causative meaning, while it implies a personal process in UNDERGO_CHANGE. The mapping, instead, links (CAUSE_CHANGE – *conversion.n*) to the *Religious_conversion* page, and all the sentences collected point to religious conversion, regardless of their causative form or not. Another characteristic of this approach is that we can acquire new lexical units regardless of their part-of-speech, even if we start from nominal lexical units. This proves that we do not need to apply the initial mapping to verbal or adjectival LUs to obtain new data for all parts of speech. For example, we linked (MEDICAL_SPECIALTIES – *gynaecology.n*) to the *Gynaecology* Wikipage. Consequently, we could include the adjective *gynaecologic*, pointing to the *Gynaecology* page, into the MEDICAL_SPECIALTIES frame for sentences like "Fellowship training in a *gynaecologic* subspeciality can range from one to four years". However, this advantage can also turn into a weakness, because *gynaecologist* is also redirected to the *Gynaecology* page, but it belongs to MEDICAL_PROFESSIONALS and should not be included into MEDICAL_SPECIALTIES.

For the 20 (F, l) pairs considered in the given sentences, it was possible also to retrieve 8 lexical units that are not present in FrameNet, for example *billy-club* for the WEAPON frame. Exploiting redirections and anchoring strategies, our induction method can account for orthographical variations, for example it acquires both *memorize* and *memorise*. On the other hand, also misspelled words may be collected, for instance *gynaecologial* instead of *gynaecological*.

7 Multilingual FrameNet expansion

One of the great advantages of Wikipedia is its availability in several languages. The English version is by far the most extended, but a considerable repository of pages is available also for other languages, esp. European ones. In general, articles on the same object in different languages are edited independently and do not have to be translations of one another, but are linked to each other by their authors. In this way, the multilingual versions of Wikipedia can be easily exploited to build comparable corpora, with connected Wikipages in different languages dealing with the same contents.

In this research step, we focus on this aspect of Wikipedia and propose a methodology that, using the English Wikipages as a bridge, automatically acquires new lexical units and example sentences also for other languages. This would represent the starting point towards the creation of FrameNet for new languages. Indeed, FrameNet structure comprises a language-independent level of information, namely frame and frame element definitions, and a language-dependent one, i.e. the lexical units and the example sentences. This makes the resource particularly suitable to corpus-based (semi) automatic creation of FrameNet for new languages, because the descriptive part can be pre-

served and the language-dependent layer can be populated with new instances in other languages (Crespo and Buitelaar, 2008).

We apply our extraction algorithm to the Italian Wikipedia. Since several approaches have been experimented to (semi) automatically build Italian FrameNet using WordNet (De Cao et al. (2008) and Tonelli and Pighin (2009)), we believe that our new proposal to exploit Wikipedia may be of interest in the research community. Anyhow, the approach can be exploited in principle for every language available in Wikipedia.

7.1 Experimental setup

Similarly to the data extraction process described in Section 6, we consider for every (F, l) pair in English the linked Wikipedia sense s, in English as well. Then, we retrieve the Italian Wikipedia sense s_i linked to s and extract all sentences C_i in the Italian Wikipedia dump[7] with a reference to s_i. In this way, we can assume that C_i are example sentences of F and that the words or expressions W_i in C_i containing an embedded reference to s_i are good candidate lexical units of F in the Italian FrameNet. For example, if we link `http://en.wikipedia.org/wiki/Court` to the JUDICIAL_BODY frame, we first retrieve the Italian version of the site `http://it.wikipedia.org/wiki/Tribunale`. Then, with a top-down strategy, we further extract all Italian sentences pointing to the `Tribunale` page and acquire as lexical units all words with an embedded reference to this concept, for example *tribunale* and *corte*. In this way, we can include the extracted lexical units and the sentences where they occur in the JUDICIAL_BODY frame for Italian.

Given the 893 (F, l) pairs in English and the linked Wikipedia senses described in 6.2, we first extracted the Italian Wikipages that are linked to the English ones. Then for every linked Wikipage in Italian, we retrieved all sentences with a reference pointing to that page in the Italian Wikipedia dump. Statistics about the extracted data are reported in Table 3.

Since the Italian Wikipedia is about one fifth of the English one, it was not possible to map every English Wikipage with an Italian article. In fact, only 371 senses out of 893 in English were linked to an Italian page. Also the average number of sentences extracted for every sense is much smaller (62 vs. 1,080). Anyhow, this does not represent a problem because in the English FrameNet, the lexical units whose annotation is considered to be *complete* are usually instantiated by set of 20 annotated sentences on average. So, according to the FrameNet standard, 60 sentences are more than enough to represent the meaning of a lexical unit in a frame.

	Italian Wikipedia
Linked Wikipages in Italian	371
N. of extracted sents	23,078
Avg. sents per Italian sense	62

Table 3: Extracted data from Italian Wikipedia

7.2 Evaluation

In this evaluation part, we took into account 1,000 sentences, in order to have a comparable dataset w.r.t. the evaluation for English. However, the sets of Italian sentences extracted for every (F, l), i.e. for every Wikipedia article, were much smaller, so we increased the number of randomly chosen (F, l) pairs to 80. Our evaluation is focused on the quality of the sentences and aims at assessing if the given sentences are correctly assigned to the (F, l) pairs. We report 69% accuracy, which is 9% lower than for English. Apart from the same errors and issues reported for English, a decrease in performance can be explained by the fact that, since less articles are present w.r.t. the English version, redirections and internal links tend to be less precise and fine-grained. For example, the word "*diritti*" in the sense of "*(human) rights*" redirects to the article about *Diritto*, corresponding to *Law* as a system of rules. On the contrary, *Law* and *Rights* have two different pages in English. Besides, the different quality of the two resources can also depend on the smaller number of users that edit and check the Italian articles. From the 1,000 sentences evaluated we extracted 145 new lexical units: since Italian FrameNet does not exist yet, every lexical unit in a sentence that is correct can be straightforwardly included in the first version of the resource.

8 Conclusions and Future work

In this work, we have proposed to apply a word sense disambiguation system to a new task, namely the linking between FrameNet and Wikipedia. Results are promising and show that

[7] `http://download.wikimedia.org/itwiki/20090203`

the task is adequately substantiated. The proposed approach can help enriching FrameNet with new example sentences and lexical units and provide a starting point for the creation of FrameNet-like resources in all Wikipedia languages. On the one hand, the retrieved data could speed up human annotation, requiring only a manual validation. On the other hand, the extracted sentences could provide enough training data to machine learning systems for frame assignment, since insufficient frame attestations in the FrameNet database are a major problem for such systems.

In the next research step, we plan to carry out an extended evaluation process in order to compute inter-annotator agreement and eventually point out validation problems. Then, we want to extend the mapping and the data extraction process to all (F, l) pairs in FrameNet (about 10,000). The retrieved sentences will be made available as training or annotation material. Besides, we want to create an online resource where the links between (F, l) pairs and Wikipages are made explicit and where users can browse the retrieved sentences. The resource can be produced and made available with a reduced effort for every language in Wikipedia. Anyway, the English version has proved to be more precise, while the resource for new languages would require a more accurate revision.

Acknowledgments

Claudio Giuliano is supported by the ITCH project (`http://itch.fbk.eu`), sponsored by the Italian Ministry of University and Research and by the Autonomous Province of Trento and the X-Media project (`http://www.x-media-project.org`), sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978.

References

Collin F. Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 Task 10: Frame Semantic Structure Extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, CZ, June.

Aljoscha Burchardt and Anette Frank. 2006. Approximating Textual Entailment with LFG and FrameNet Frames. In *Proceedings of the 2nd PASCAL RTE Workshop*, pages 92–97, Venice, Italy.

Diego De Cao, Danilo Croce, Marco Pennacchiotti, and Roberto Basili. 2008. Combining Word Sense and Usage for modeling Frame Semantics. In *Proceedings of STEP 2008*, Venice, Italy.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. 2006. *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Ian Chow and Jonathan Webster. 2007. Integration of Linguistic Resources for Verb Classification: FrameNet Frame, WordNet Verb and Suggested Upper Merged Ontology. *Computational Linguistics and Intelligent Text Processing*, pages 1–11.

Mario Crespo and Paul Buitelaar. 2008. Domain-specific English-to-Spanish Translation of FrameNet. In *Proc. of LREC 2008*, Marrakech.

Andras Csomai and Rada Mihalcea. 2008. Linking Documents to Encyclopedic Knowledge. *IEEE Intelligent Systems, special issue on "Natural Language Processing for the Web"*.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June. Association for Computational Linguistics.

Katrin Erk. 2004. Frame assignment as Word Sense Disambiguation. In *Proceedings of IWCS-6*, Tilburg, NL.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

C.J. Fillmore, C.R. Johnson, and M. R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250, September.

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual ACL meeting*, pages 929–936, Morristown, US.

A. Gliozzo, C. Giuliano, and C. Strapparava. 2005. Domain kernels for word sense disambiguation. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL-05)*, pages 403–410, Ann Arbor, Michigan, June.

R. Johansson and P. Nugues. 2007. Using WordNet to extend FrameNet coverage. In *Proc. of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*, Tartu.

B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. 2002. The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering*, 8(4):359–373.

R. Mihalcea and P. Edmonds, editors. 2004. *Proceedings of SENSEVAL-3*, Barcelona, Spain, July.

David Milne and Ian H. Witten. 2008. Learning to link with Wikipedia. In *CIKM '08: Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518, NY, USA. ACM.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June. Association for Computational Linguistics.

J. Shawe-Taylor and N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Lei Shi and Rada Mihalcea. 2004. Open Text Semantic Parsing Using FrameNet and WordNet. In *Proceedings of HLT-NAACL 2004*.

Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Proceedings of CICLing 2005*, pages 100–111. Springer.

Sara Tonelli and Daniele Pighin. 2009. New features for FrameNet - WordNet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Boulder, CO, USA.