Artificial Intelligence 194 (2013) 203–221

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Wikipedia-based WSD for multilingual frame annotation

Sara Tonelli [*], Claudio Giuliano, Kateryna Tymoshenko

Fondazione Bruno Kessler, via Sommarive 18, I-38100 Trento, Italy

ARTICLE INFO

Article history:
Received 11 November 2010
Received in revised form 9 May 2012
Accepted 16 June 2012
Available online 18 June 2012

Keywords:
Frame annotation
Multilingual FrameNets
Word sense disambiguation
FrameNet–Wikipedia mapping

ABSTRACT

Many applications in the context of natural language processing have been proven to achieve a significant performance when exploiting semantic information extracted from high-quality annotated resources. However, the practical use of such resources is often biased by their limited coverage. Furthermore, they are generally available only for English and few other languages.

We propose a novel methodology that, starting from the mapping between FrameNet lexical units and Wikipedia pages, automatically leverages from Wikipedia new lexical units and example sentences. The goal is to build a reference data set for the semi-automatic development of new FrameNets. In addition, this methodology can be adapted to perform frame identification in any language available in Wikipedia.

Our approach relies on a state-of-the-art word sense disambiguation system that is first trained on English Wikipedia to assign a page to the lexical units in a frame. Then, this mapping is further exploited to perform frame identification in English or in any other language available in Wikipedia. Our approach shows a high potential in multilingual settings, because it can be applied to languages for which other lexical resources such as WordNet or thesauri are not available.

1. Introduction

The FrameNet database [1,2] is an English lexical resource based on the description of some prototypical situations, the *frames*, and the frame-evoking words or expressions associated to them, the *lexical units*. Every frame corresponds to a scenario involving a set of participants, the *frame elements*, that are typically the semantic arguments shared by all lexical units in a frame. Given the rich semantic information provided by frames, there have been several attempts to exploit this knowledge to improve diverse natural language processing (NLP) tasks, from question answering [3] to relation extraction [4], and entailment rules generation [5]. The integration of this semantic paradigm in existing NLP tools, however, has been hindered by difficulties in creating systems for frame semantic parsing. Some attempts have been made, using FrameNet data for training [6–8]. Since large amounts of data with high-quality annotation are currently available only in English [1] and German [9], however, the applicability of supervised approaches has been limited to these two languages. Alternative approaches based on systems that are not trained directly of FrameNet have been only partially explored by investigating the integration between FrameNet and WordNet [10–12] and the use of distributional approaches [13,14].

In this article, Wikipedia is used as an extensive, multilingual repository of frame information in order to achieve two main goals: first, to devise a novel approach to multilingual *frame identification*, a subtask of frame semantic parsing, without training a system directly on FrameNet. Then, to retrieve a large amount of *frame example sentences* in different languages.

[*] Corresponding author. Tel.: +39 0461 314 542; fax: +39 0461 314 591.
E-mail addresses: satonelli@fbk.eu (S. Tonelli), giuliano@fbk.eu (C. Giuliano), tymoshenko@fbk.eu (K. Tymoshenko).

0004-3702/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artint.2012.06.002

We rely on Wikipedia because of several reasons. First of all, it is the largest existing repository of encyclopedic knowledge, freely available in 282 languages. It combines free-form natural language content with structural information, represented by intra- and inter-language links. Furthermore, it generally shows high editorial quality, especially the Wikipedia versions of widely used languages.

In order to exploit Wikipedia to perform frame annotation, a strategy to link this resource to FrameNet has been proposed. However, we are aware that inter-operability between FrameNet and Wikipedia may be hindered by the different structure, granularity and extension of the two resources. Therefore, three main research questions are addressed in this article: (i) Is it possible to link FrameNet and Wikipedia and to exploit the outcome of this mapping for frame identification? (ii) Which strategy can be chosen to devise a frame identification system that is not directly trained on FrameNet examples? And how does it compare with state-of-the-art systems? (iii) Can the same strategy be employed to support the development of non-English FrameNets? To which extent?

The first question has been partially addressed in the preliminary work by Tonelli and Giuliano [15], in which the idea to use Wikipedia as a multilingual repository of frame information was first presented. The second problem, instead, has not been tackled before. We address it by comparing our approach with a state-of-the-art frame semantic parser for English. As for multilingual frame annotation, the acquisition of frame example sentences from Italian Wikipedia was introduced by [15], although it was only marginally evaluated. On the contrary, a methodology to use the same frame identification approach for different languages is presented for the first time in this article, and is evaluated by comparing it with WordNet-based strategies [11].

This article is structured as follows. We introduce FrameNet and Wikipedia in Section 2. We present past research work related to our approach in Section 3. A general description of our methodology is provided in Section 4. The Wikipedia-based disambiguation system is described and compared with the state of the art in Section 5. The methodology for mapping frame–lexical unit pairs with Wikipedia pages is described and evaluated in Section 6, in which the first of our research questions is addressed (see items above). Then, in Section 7 a new frame identification approach is described and applied to English lexical units. A thorough evaluation and a comparison with the state-of-the-art SEMAFOR system [8] are reported, addressing our second research topic. Section 8 is devoted to our third research question and details a two-fold strategy for the creation of multilingual FrameNets: first, example sentences and lexical units in a new language are extracted from Wikipedia, and then the word sense disambiguation (WSD) system is used for multilingual frame identification. Finally, we draw some conclusions and discuss future work in Section 9.

2. FrameNet and Wikipedia: description and terminology

FrameNet [1,2] is a lexical resource for English, based on frame semantics [16], that is being created in the context of the *Berkeley FrameNet project*.[1] Its aim is to collect the range of semantic and syntactic combinatorial possibilities of each word in each of its senses through the annotation of example sentences. The conceptual model is based on three main elements:

- **Semantic frames**: Cognitive schemata or scenarios necessary to understand the meaning of words. They describe situations, objects and events and the participants involved in them.
- **Lexical units (LUs)**: Words, multiwords, idiomatic expressions evoking a frame.
- **Frame elements (FEs)**: Semantic roles involved in the situation or event expressed by a frame. They apply to all LUs in the same frame.

FrameNet 1.3, released in 2006, is comprised of more than 10,195 lexical units, 6000 of which are fully annotated, and nearly 800 semantic frames with hierarchical relations. An essential element of the FrameNet database is the *corpus-based evidence*, i.e., every lexical has to be instantiated by at least one example sentence. In FrameNet 1.3, more than 135,000 sentences have been manually annotated with frame information.

As an example, we report in Table 1 the FrameNet entry for the Wearing frame.

In the first row, the frame definition in natural language is reported, while the second includes the list of the *core* frame elements. The third row contains part of the LU list including all frame-evoking predicates, while in the fourth a few example sentences are reported. All LUs are printed in bold, while the phrases bearing a FE label are reported between square brackets, followed by the role label.

In the remainder of this article, we call *frame semantic annotation* the annotation of sentences with both frame and FE (or role) information, as performed by *frame-semantic parsers* (e.g. [6] and [8]). The sub-task of assigning a frame label to a lexical unit in a sentence is called *frame identification*. This concerns both lexical units that are listed in FrameNet, the so-called *seen LUs*, and those that are not present in the resource, the *unseen LUs*. When frame identification is applied to unseen LUs, and leads to the acquisition of new LUs, it is also known as *LU induction* [13].

The second resource we take into account in this work is *Wikipedia*, the largest online repository of encyclopedic knowledge. At the moment of writing, there are 20 million articles in 282 languages (over 3.82 million in English alone) written

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

Table 1
Wearing frame.

	Frame: Wearing	
Def.	The words in this frame refer to what clothing a wearer (or a specific body_part of the wearer) has on.	
FEs	body_part	The body part of the wearer which is covered by the clothing.
	clothing	This FE identifies the clothing that the wearer wears.
	wearer	The person whose clothing is under discussion.
LUs	attired.a, bare-armed.a, bare-breasted.a, bare.a, braless.a, clothed.a, coatless.a, costumed.a, decked out.a, dressed.a, have got on.v, sport.v, swaddled.a, swathed.a, wear.v [...]	
Ex.	[The leader]$_{\text{Wearer}}$ **wore** [a golden helmet]$_{\text{Clothing}}$. She saw that [her]$_{\text{Wearer}}$ [left hand]$_{\text{Body\_part}}$ was **bare**. [She]$_{\text{Wearer}}$ **had** [an apron]$_{\text{Clothing}}$ **on**.	

collaboratively by approximately 100,000 regularly active contributors around the world. This makes Wikipedia a reliable source of knowledge both for Internet users and researchers.

The *article* (or *page*) is the basic entry in Wikipedia. Each article is uniquely identified by an 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 many links or *anchors* that connect the most relevant terms to other pages. Such connections are manually added by Wikipedia contributors following the available *Manual of Style*[2] and are used to increase the reader's understanding of the topic and to find related information.

3. Related work

Frames [17] are primarily *cognitive structures* determined by the social environment and personal experiences of an individual rather than by the language. This assumption was first analyzed at lexical level by Boas [18], who considered semantic frames as inter-lingual representations for multilingual lexical databases. The language-independence hypothesis has been to a great extent confirmed by many projects aimed at the development of FrameNet in different languages starting from a common repository of English-based frames [19–23]. These projects showed that only in few cases the English-based model cannot capture relevant semantic distinctions that are specific to another language [24]. Therefore, in this article we make the crucial assumption that frames are constant across languages.

FrameNet has often been mapped to WordNet [25], mainly for LU induction, with the aim of improving FrameNet data sparseness. Different approaches have been proposed, which aim at first assigning a WordNet synset to a LU through a WSD step and then at linking the given synset to a frame by exploiting synonym and hypernym information in WordNet (see for example the rule-based approach by Burchardt et al. [10] and the SVM-based techniques presented in [26,27,13,11]). Other authors have proposed to exploit semi-automatic mappings between WordNet, FrameNet and other resources, for example VerbNet [12] and LDOCE [28].

The first task we address in this work is frame identification on English documents, with a focus on unseen lexical units. Current systems for frame annotation [7,6,8] are usually trained on FrameNet examples in order to assign the most probable frame to a candidate lexical unit. In case a lexical unit is not present in FrameNet, WordNet is usually exploited as a detour to assess similarity between seen and unseen lexical units [10]. Supervised systems integrate WordNet-based features to tackle this problem [8,26]. Some other attempts have been made to integrate WordNet-based and distributional models [13]. We address the problem from a completely new perspective, in that we propose to base the disambiguation step on Wikipedia sense repository rather than on FrameNet frames. In contrast to previous works, this approach is also applicable to many different languages.

As for the *induction of LUs and example sentences in new languages*, which is an issue tackled in Section 8 of the present work, previous attempts have been made using different lexical resources. In most cases, WordNet was first mapped with FrameNet, then its multilingual extensions were used for LU induction in different languages. For example, see the works by de Cao et al. [27] and Tonelli and Pighin [11] based on MultiWordNet [29] for the acquisition of Italian LUs, and Crespo and Buitelaar [30] exploiting *EuroWordNet* [31] for retrieving Spanish LUs. Other approaches that do not rely on WordNet have been proposed for French [32], Chinese [33] and German [34]. In the first case, new French LUs were acquired by translating English lexical units with the *Wiktionary* and the *EuRADic* dictionary. As for Chinese, the LUs in the Berkeley FrameNet were first mapped with entries listed in the *HowNet* Chinese ontology. Then, some sentences containing the mapped predicates were extracted from a Chinese corpus and filtered according to specific PoS-sequences. The methodology presented for German, instead, does not make use of bilingual dictionaries and only relies on automatically aligned parallel corpora and a set of filters for translation inconsistencies.

[2] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style.

In contrast, our approach to the acquisition of LUs and example sentences in new languages relies neither on bilingual dictionaries nor on aligned parallel corpora. Also, we do not carry out any linguistic analysis of the sentences before extracting them. The general framework resembles the one of WordNet-based approaches, in that we also start from a mapping between FrameNet and another English resource (in our case Wikipedia), and then exploit the multilingual potential of such resource to acquire data in other languages. The advantage of using Wikipedia instead of WordNet-related resources is that many more languages are represented, it is continuously updated, and semi-structured information is coupled with free-form natural language texts.

With regard to Wikipedia, two research directions are closely related to our work. The first aims at using Wikipedia content in order to extend existing resources with limited coverage or to create new ones. The second concerns the task of automatically enriching plain text with links to Wikipedia. With respect to the first research field, Wikipedia has been combined and mapped to several resources, with WordNet being the most widely used. Ruiz-Casado et al. [35], for example, propose a methodology to map WordNet synsets to the articles of *Simple Wikipedia* (SW), which is a version of Wikipedia using simple English words and grammar.[3]

Medelyan and Legg [36] integrate *Cyc* [37] knowledge base with Wikipedia by mapping Wikipedia articles to Cyc concepts, with the purpose of further extending Cyc with Wikipedia knowledge such as new synonyms and translations. Given a Cyc concept, a set of potential Wikipedia mappings is found by exploiting synonymy information from both resources. If the set contains more than one possible mapping, the best Wikipedia page is chosen based on its commonness and semantic relatedness to the context of a Cyc concept. Such context is created using the categories that surround the concept in the Cyc taxonomy, e.g. direct hypernyms and hyponyms. More recently, the mapping algorithm has been further improved by Sarjant et al. [38], and Wikipedia has been used to create new child concepts for Cyc categories.

Suchanek et al. [39] use Wikipedia and WordNet to automatically construct a knowledge base called *YAGO*.[4] YAGO classes are derived from WordNet taxonomy, while Wikipedia category system is used to further extend the YAGO class taxonomy. The extension algorithm is based on parsing the category names and mapping their constituents to WordNet using the most frequent sense strategy. Wikipedia pages are used to populate the classes with individuals. Other resources built with the use of Wikipedia include, among others, DBpedia[5] and Freebase.[6] Navigli and Ponzetto [40] create the large-scale multilingual resource *BabelNet* by combining Wikipedia and WordNet. The authors perform disambiguation when necessary, and make the resource multilingual by using Wikipedia cross-language links and machine translation techniques.

The other line of research on Wikipedia related to our work is the automatic annotation of terms in a plain text with links to Wikipedia pages. This is a WSD task because its goal is to link a term in a sentence to the Wikipedia concept that best expresses its sense. Some well-known approaches to this task include the works by Csomai and Mihalcea [41] and by Milne and Witten [42]. They perform the 'wikification' of the document, that is they identify the main concepts in a text and annotate them with links to Wikipedia pages. Csomai and Mihalcea [41] divide the task into two steps, namely the extraction of relevant concepts and the WSD step, with Wikipedia pages as a sense repository. The second step is closely related to our work. In this step, the authors experiment with a knowledge-based WSD algorithm and a data-driven one. The second approach, which integrates local and topical features into a naive-Bayes classifier, achieves better results. The methodology that we are proposing in this paper is similar in spirit, but it uses more sophisticated features and machine learning techniques. Milne and Witten [42] also decompose the task into two steps, but in reversed order. First, all terms in a document are possibly linked to appropriate Wikipedia pages, then the most relevant links are selected. The pages to which terms in the text can be linked unambiguously form the *context*. Then disambiguation of a specific term is performed through a machine-learning approach, using the commonness of each sense in Wikipedia, its relatedness to the context, and the context coherence as features. The approach achieves competitive results compared to [41]. However, its limitation is that it relies on presence of non-ambiguous terms in the document, which is not always the case.

Kulkarni et al. [43] address a more general task, aiming at exhaustive annotation of a document with links to Wikipedia. Based on the assumption that entities in the same document tend to be topically related, they cast the task as a collective optimization problem. They annotate a document with links to Wikipedia by maximizing a function which encodes the joint annotation probability. It incorporates compatibility between terms contexts and candidate Wikipedia pages, and topical coherence of all the Wikipedia pages to which terms in the document have been linked.

4. General workflow for multilingual frame annotation

The core component of our methodology for multilingual frame identification and acquisition of frame example sentences is a word sense disambiguation system based on Wikipedia. It relies on a supervised approach that, given training examples extracted from Wikipedia, assigns the correct sense (a Wikipedia article) to a specific term in a text.

The system is part of a workflow based on five steps. The *first* is the creation of a sense-tagged training set based on the English Wikipedia and, then, the training of the WSD system (Section 5).

[3] http://simple.wikipedia.org/.

[4] http://www.mpi-inf.mpg.de/yago-naga/yago/.

[5] http://dbpedia.org/.

[6] http://www.freebase.com/.

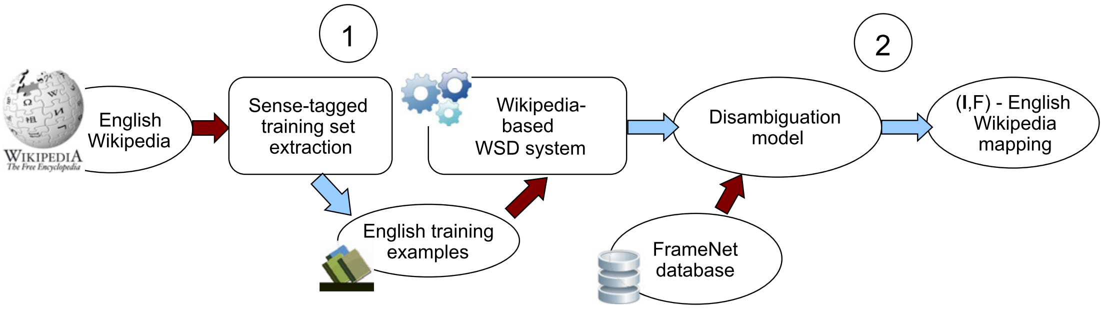
1
2
English Wikipedia
WIKIPEDIA The Free Encyclopedia
Sense-tagged training set extraction
Wikipedia-based WSD system
Disambiguation model
(l,F) - English Wikipedia mapping
English training examples
FrameNet database

Fig. 1. Creation of a WSD model based on Wikipedia and mapping with FrameNet.

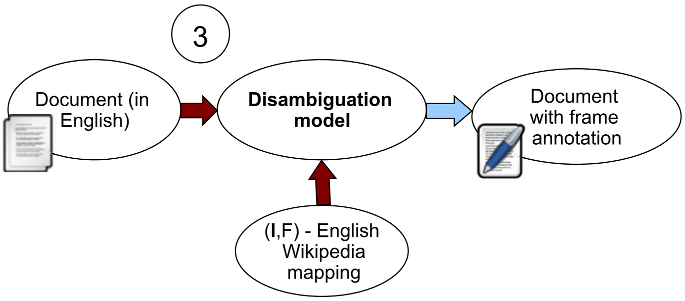
3
Document (in English)
Disambiguation model
Document with frame annotation
(l,F) - English Wikipedia mapping

Fig. 2. Frame identification on English documents.

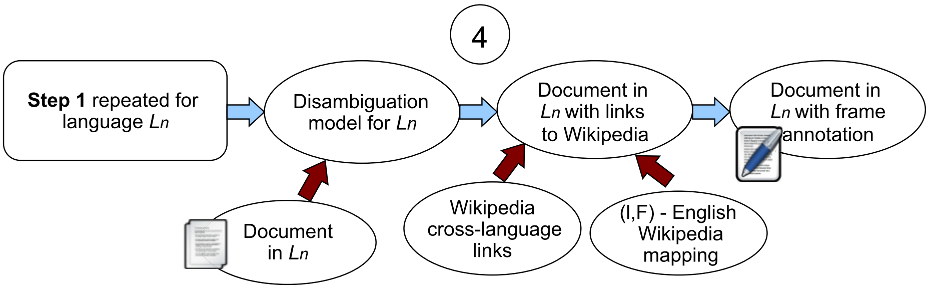
4
Step 1 repeated for language Ln
Disambiguation model for Ln
Document in Ln with links to Wikipedia
Document in Ln with frame annotation
Document in Ln
Wikipedia cross-language links
(l,F) - English Wikipedia mapping

Fig. 3. Frame identification for language L_n.

In the *second step*, the English WSD system is used to create a mapping between English Wikipedia articles and (l, F) pairs, with l being a lexical unit in a frame F defined in FrameNet (Section 6). The first two steps are presented in Fig. 1.

In the *third* step, the English WSD system and the mapping between FrameNet and Wikipedia are exploited for frame identification over seen and unseen lexical units in English texts. The WSD system is first used to disambiguate all words in a text and, then, the mapping is used to associate them with a frame label (Section 7). This step is portrayed in Fig. 2.

In the *fourth* step, the WSD system is trained on Wikipedia in a non-English language L_n. It is then used to disambiguate texts in L_n by linking each term to a Wikipedia article in L_n. Next, the cross-language links between articles in L_n and English are exploited to retrieve the English version of the linked article. Finally, the mapping between English Wikipedia and FrameNet is applied again to annotate the texts in L_n with frame labels (Section 8.2). This step is depicted in Fig. 3.

In addition, frame example sentences are acquired for the language L_n from the corresponding Wikipedia version. We take each English Wikipage w_e previously mapped to a frame f and extract its version w_n in the language L_n. Then, we retrieve from Wikipedia all sentences pointing to w_n, and we acquire them as example sentences of the frame f in the new language L_n. All terms in the sentences that point to w_n are retrieved as lexical units of f (Section 8.1).

5. The WSD system

In the proposed framework, the two problems of mapping LUs to Wikipedia articles and identifying frames in multilingual documents are cast as a WSD exercise. WSD is the task of selecting the appropriate sense of a word in a text, according to a sense inventory containing a list of possible senses for each word [44,45]. Our WSD system, called *The Wiki Machine*,[7] is an ensemble of word-expert classifiers trained on labeled data extracted from Wikipedia annotations. The following sections present the machine learning approach and its evaluation.

5.1. Learning algorithm

We have extended the kernel-based approach described by Giuliano et al. [46], in which basic kernel functions are employed to integrate syntactic, semantic, and pragmatic knowledge sources typically used in the WSD literature [44,45]. Kernel methods are theoretically well founded in statistical learning theory and show good empirical results in many applications [47]. The strategy adopted consists in splitting the learning problem into two parts. First, the input data are embedded in a suitable feature space, and then a linear algorithm (in our case, support vector machines) is used to discover nonlinear patterns in the input space. The kernel function is the only task-specific component of the learning algorithm. By exploiting the properties of kernels, basic kernels are then combined to define the WSD kernel. Specifically, we used a linear combination of local and global kernels defined in the following sections.

5.1.1. Local kernels

The local kernel extends the gap-weighted subsequence kernel to capture syntagmatic relations between words. Typically, n-grams of words and part-of-speech tags extracted from the local context of the target word are used to represent syntagmatic relations [48]. However, n-grams fail to represent non-contiguous or shifted collocations, and to consider lexical variability. For example, suppose we have to disambiguate the verb *to score* in the context "Maradona scored Argentina's third goal", given the labeled example "Ronaldo scored two goals in the second half" as training. A traditional approach, that only considers contiguous n-grams, has no clues to conclude that the two occurrences of the verb have the same sense because the two contexts have no features in common. Instead, the gap-weighted subsequence kernel can extract the non-contiguous bigram "score goal", shared by the two examples. Sequence kernels are a family of kernel functions developed to compute the inner product among images of strings in a high-dimensional feature space using dynamic programming techniques [49–51]. The gap-weighted subsequence kernel is one of the most general types of kernel functions based on sequences (aka string kernels). Roughly speaking, it compares two strings by means of the number of contiguous and non-contiguous substrings of a given length they have in common. Non-contiguous occurrences are penalized according to the number of gaps they contain.

Formally, let V be the vocabulary, the feature space associated with the sequence kernel of length n is indexed by a set I of subsequences over V of length n. The (explicit) mapping function is defined by

$$\phi_u^n(s) = \sum_{\mathbf{i}:\ u=s(\mathbf{i})} \lambda^{l(\mathbf{i})}, \quad u \in V^n, \tag{1}$$

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 kernel function is defined by

$$K_n(s_1, s_2) = \langle \phi_u^n(s_1), \phi_u^n(s_2) \rangle = \sum_{u \in V^n} \phi_u^n(s_1)\phi_u^n(s_2). \tag{2}$$

Furthermore, the kernel defined in Eq. (2) is extended to compare all subsequences of length up to p. Formally,

$$K_p(s_1, s_2) = \sum_{n=1}^{p} K_n(s_1, s_2). \tag{3}$$

The local kernel is obtained as a combination of extended gap-weighted subsequence kernels defined on sequences of word stems, part-of-speech tags, and some orthographic features extracted from a fixed-size window centered on the target word. This implementation differs from the original in the removal of the soft-matching criteria and in the introduction of some orthographic features. The first change is due to a significant increase of computational cost compared with the performance improvement we can obtain, since Giuliano et al. [46] report +0.8% on the English lexical sample task of SemEval 2007. The introduction of orthographic features, typically employed in named entity recognition, helps in distinguishing between nouns and name senses present in the sense inventory derived from Wikipedia, without appreciably affecting the computational effort. The integration of orthographic features is straightforward, as we have simply to define a function that takes as input a word and returns *CAP*, *UPPER*, *LOWER*, *ALPHANUM*, *PUNCT*, *NUM*, or *SYMB* if the word is capitalized, in uppercase, in lowercase, a sequence of alpha-numeric characters, a punctuation, a numeral, or a symbol respectively.

Formally, the local kernel is defined by

$$K_L(s_1, s_2) = S_p(s_1, s_2) + P_p(s_1, s_2) + O_p(s_1, s_2), \tag{4}$$

[7] http://thewikimachine.fbk.eu/.

where S_p, P_p, and O_p are extended gap-weighted subsequences (Eq. (3)) defined on sequences of word stems, part-of-speech tags and the three orthographic features defined above, respectively. It follows directly from the closure properties of kernels that it is a valid kernel.

5.1.2. Global kernels

The global kernel combines the bag-of-words and latent semantic kernels to capture semantic, domain, and topical information. This composite kernel takes as input a wide context window around the target word.

Specifically, the bag-of-words kernel defines an N-dimensional feature space, in which the context c is represented by a row vector

$$\phi(c) = \big(tf(t_1,c), tf(t_2,c), \ldots, tf(t_N,c)\big) \in \mathbb{R}^N, \tag{5}$$

where the ith component is the frequency of the word t_i in c. The bag-of-words kernel is defined as

$$K_{BOW}(c_1,c_2) = \sum_{j=1}^{N} tf(t_j,c_1)tf(t_j,c_2). \tag{6}$$

The main drawback of this approach is the need of a large amount of training data to reliably generalize over unseen data. For example, despite the fact that the examples “People affected by AIDS” and “HIV is a virus” express related concepts, their similarity is zero using the bag-of-words model since they have no words in common (they are represented by orthogonal vectors in the vector space model). On the other hand, due to the ambiguity of the word *virus*, the similarity between the contexts “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 the drawback of bag-of-words, we incorporate semantic information by means of latent semantic kernel [52]. The contexts are implicitly mapped into a “semantic space” where documents that do not share any words can still be close to each other if their words are semantically related. The semantic similarity model is obtained by the co-occurrence analysis of a large corpus: words that co-occur often in the same documents are considered related. The technique used to extract the co-occurrence statistics relies on a singular value decomposition (SVD) of the term by the document matrix of the corpus. The contextual features are projected into the subspace spanned by the first k singular vectors of the feature space. Thus, the dimension of the feature space is reduced to k and its dimension can be controlled by varying k. For example, the similarity in the latent semantic space of the two examples “People affected by AIDS” and “HIV is a virus” is higher than in the bag-of-words representation, because the terms “AIDS”, “HIV”, and “virus” very often co-occur in the medical domain.

Formally, the matrix $\mathbf{D}$ is used to define a function $\mathcal{D}: \mathbb{R}^N \to \mathbb{R}^k$, that maps the vector $\phi(c)$ represented in the standard bag-of-words space, into the vector $\phi'(c)$ in the latent semantic space. $\mathcal{D}$ is defined as

$$\mathcal{D}\big(\phi(c)\big) = \phi(c)\big(\mathbf{I^{IDF}D}\big) = \phi'(c), \tag{7}$$

where $\mathbf{I^{IDF}}$ is an $N \times N$ diagonal matrix such that $i^{IDF}_{i,i} = IDF(t_i)$, and $IDF(t_i)$ is the inverse document frequency of t_i.

SVD is used to obtain the matrix $\mathbf{D}$ from a corpus represented by its term-by-document matrix $\mathbf{T}$. SVD decomposes the matrix $\mathbf{T}$ into three matrices $\mathbf{T} \simeq \mathbf{V\Sigma_k U}^T$, 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 $\mathbf{TT}^T$ and $\mathbf{T}^T\mathbf{T}$ respectively, and $\mathbf{\Sigma_k}$ is the diagonal $N \times N$ matrix containing the highest $k \ll N$ eigenvalues of $\mathbf{T}$, and all the remaining elements set to 0. The parameter k is the dimensionality of the latent semantic space and can be fixed in advance. Under this setting, the matrix $\mathbf{D}$ is defined as

$$\mathbf{D} = \mathbf{I^N V}\sqrt{\mathbf{\Sigma_k}} \tag{8}$$

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 ith row of the matrix $\mathbf{V}\sqrt{\mathbf{\Sigma_k}}$. The latent semantic kernel is explicitly defined as

$$K_{LS}(c_1,c_2) = \big\langle \mathcal{D}(c_1), \mathcal{D}(c_2) \big\rangle, \tag{9}$$

where $\mathcal{D}$ is the mapping defined in Eq. (7).

5.1.3. Composite kernel

Finally, to combine local and global information, the composite kernel is defined by

$$K_{WSD}(t_1,t_2) = \hat{K}_L(t_1,t_2) + \hat{K}_{BOW}(t_1,t_2) + \hat{K}_{LS}(t_1,t_2), \tag{10}$$

where $\hat{K}_L$, $\hat{K}_{BOW}$, and $\hat{K}_{LS}$ are normalized kernels defined in Eqs. (4), (6), and (9), respectively.[8] It follows directly from the explicit construction of the feature space and from closure properties of kernels that it is a valid kernel.

[8] $\hat{K}(x_1,x_2) = \frac{K(x_1,x_2)}{\sqrt{K(x_1,x_1)K(x_2,x_2)}}$.

5.1.4. Training set

The labeled examples are extracted from Wikipedia annotations, as first proposed by Mihalcea [53]. Specifically, we parse Wikipedia, and for each internal link, we extract the anchor term, the target article, and the surrounding context. The target article is regarded as sense annotation for the corresponding term in context (the example). Examples are then grouped according to the anchor term to create the training sets for the word-expert classifiers. For instance, in the context "Lie theory is frequently built upon a study of the classical linear algebraic groups. Special branches include Weyl groups, Coxeter groups, and `[[buildings|Building_(mathematics)]].`", extracted from the Wikipedia article `Lie_theory`, we assume that the word *building* has the meaning defined by the article `Building_(mathematics)`. Overall, the word *building* is found as a link to 42 different articles in 708 different contexts. This constitutes the sense inventory and the training set to train the "building"-expert classifier. Note that links are in most cases accurate, as they are manually created by the Wikipedia contributors.

We extracted the English and Italian training sets from the May 2010, English Wikipedia dump[9] and the June 2010, Italian Wikipedia dump,[10] respectively, and the dictionaries have 12,321,704 and 2,369,918 different entries, that correspond to 101,105,787 and 14,429,138 labeled examples. Links to disambiguation pages are not considered and links to redirection pages are replaced with the redirected page.

5.2. Implementation details

We used the following open source tools in our implementation of the algorithm: the Java Wikipedia Library to parse the Wikipedia dumps [54]. The Apache OpenNLP library for sentence detection, tokenization, and part-of-speech tagging.[11] The Snowball library for stemming.[12] The LIBSVM library for support vector machines [55]. The SVDLIBC package to compute the SVD.[13]

The following set up is used in our experiments. The English and Italian latent semantic models are derived from the 200,000 most visited Wikipedia articles in each language.[14] After removing terms that occur less than 5 times, the resulting dictionaries contain about 300,000 terms. The decomposed matrix D is truncated to 100 dimensions (parameter k). No task-specific parameter optimization was performed during the experiments. We used the default LIBSVM parameter settings. The global context corresponds to a paragraph and the local context to a window of three words before and after the target word. The local kernel parameter p is set to 3.

5.3. System evaluation

The original algorithm [46] achieved state-of-the-art results for a wide range of languages at Senseval-3 [56] and SemEval 2007 [57] evaluation exercises. In addition, to provide an up-to-date evaluation, we assessed the system on the ACE05-WIKI Extension [58]. This benchmark extends the English Automatic Content Extraction (ACE) 2005 data set with ground-truth links to Wikipedia.[15] It is specifically designed to evaluate disambiguation systems based on Wikipedia. ACE 2005 is composed of 599 articles assembled from a variety of sources, selected from broadcast news programs, newspapers, newswire reports, Internet sources, and transcribed audio. It contains the annotation of entity mentions of different types, such as person, location, and organization. In the extension, each name (NAM) and nominal (NOM) mention (in total 29,300 entity mentions) is manually assigned to zero or more Wikipedia articles. If assigned to more than one, they are ordered from most specific to most generic.

We have compared our approach to the state-of-the-art system *Wikipedia Miner* [42]. Since the system requires that Wikipedia is preprocessed in a specific way, we used the preprocessed version of July 2008, made available by the authors. Table 2 shows the results obtained on the ACE05-WIKI Extension. The evaluation is performed considering only the most specific Wikipedia articles assigned by human annotators as gold standard annotations. We report the evaluation of the full task, consisting in linking both name and nominal mentions (NAM & NOM), and the partial tasks, namely the linking of name mentions (NAM) and nominal mentions (NOM). The latter evaluation is more interesting for our purposes, since lexical units do not include proper names.

The Wiki Machine significantly outperforms Wikipedia Miner in the NAM & NOM and NOM tasks. The difference between the two systems is due to the smaller number of knowledge sources exploited by Wikipedia Miner. Specifically, Wikipedia Miner does not use the Wikipedia internal links and, consequently, the information provided by their local and global contexts. This difference is further stressed in the disambiguation of nominal mentions, in which the use of information extracted from the local context is more important than in the disambiguation of name mentions. For example, in the phrase "opening goal" the fact that the word *goal* is preceded by *opening* suggests that it is used in the sense of "a successful

[9] http://download.wikimedia.org/enwiki/20100312/.

[10] http://dumps.wikimedia.org/itwiki/20100624/.

[11] http://incubator.apache.org/opennlp/.

[12] http://snowball.tartarus.org/.

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

[14] Wikipedia article traffic statistics are available at http://stats.grok.se/.

[15] http://www.itl.nist.gov/iad/mig//tests/ace/ace05/index.html.

Table 2
Comparative evaluation of the two disambiguation methods on ACE05-WIKI (micro-average). Symbol † indicates significant differences relative to the corresponding mention type ($p < 0.01$). Significance tests are computed using approximate randomization procedure.

Approach	Mention type	Precision	Recall	F_1
Wikipedia Miner	NAM & NOM	0.78†	0.48	0.59
	NAM	0.86†	0.69	0.76
	NOM	0.66	0.28	0.40
The Wiki Machine	NAM & NOM	0.72	0.71†	0.71†
	NAM	0.78	0.74†	0.76
	NOM	0.62	0.65†	0.63†

Table 3
Comparative evaluation of the basic and composite kernels on ACE05-WIKI. The use of orthographic features is denoted by *. Symbols † and ‡ indicate a significant difference with respect to the preceding entry in the row, $p < 0.01$ and $p < 0.05$, respectively. † or ‡ scores indicate significant differences relative to preceding entry in the row ($p < 0.01$ and $p < 0.05$, respectively). Significance tests are computed using approximate randomization procedure.

	Basic kernels			Composite kernels		
	K_L	K_L^*	K_{BOW}	K_{BOW+LS}	K_{WSD}	K_{WSD}^*
F_1	62.3	63.2‡	64.6‡	66.1†	69.7†	71.0‡

attempt at scoring". In contrast, the local context provides few clues to disambiguate a name entity, for example, in the phrase "Mr. Johnson" the fact that the word *Johnson* is preceded by the title *Mr.* can suggest the mention type, number, gender, etc., but only the global context provides decisive clues to identify the corresponding Wikipedia article. Finally, the two-year difference between the training data used by the two systems may partly affect their performance. To alleviate this problem, we updated the output of Wikipedia Miner, taking into account the changes occurred in the redirection pages, which led to an improvement of 1 point. On the contrary, varying the Wikipedia Miner free parameters did not produce a statistically significant improvement.

Table 3 shows the comparison between the basic and composite kernels. The results substantially confirm previous conclusions [46] that the composite WSD kernel significantly outperform the basic ones. In addition, significant differences between the kernels with and without the orthographic features show the usefulness of this additional information.

Finally, Mendes et al. [59] compared The Wiki Machine with an ensemble of academic and commercial systems, namely, DBpedia Spotlight, Zemanta, Open Calais, Alchemy API, and Ontos, showing that our system has the highest F-score.

6. FrameNet–Wikipedia mapping

We apply the WSD model learned as described in Section 5 to assign Wikipedia articles to lexical units. We do so by creating a pseudo-context for each (l, F) pair, in which the frame definition and the lexical units associated to F build the left and the right context of the lexical unit to be disambiguated. As an example, we report below the pseudo-context created for the lexical unit *cable.v* in the Communication_means frame:

> This frame concerns Communicators communicating with each other with the aid of a Means of communication such as a telephone *cable* wire, phone, semaphore, telegraph, telex, radio, telephone, fax.

The disambiguation algorithm assigns a sense, i.e. a Wikipedia article, to the lexical unit in the pseudo-context, so that such sense uniquely defines the mapping.

6.1. Mapping lexical units to Wikipedia

We restrict the mapping to nominal lexical units as Wikipedia is basically a resource organized in concepts, usually expressed by nouns. Verbs and adjectives are generally linked to articles describing nominal concepts.

We first extract from FrameNet all (l, F) pairs with a nominal l, obtaining a set of 4154 lexical units. Then, the disambiguation step is performed based on each pseudo-context. Table 4 shows the mapping statistics. We compare the disambiguation step with an informed *baseline* that assigns to l the page to which it is most frequently linked in Wikipedia. For example, the baseline sense for the pair (*bonnet.n*, Accoutrements) is the Wikipedia article {Hood_(vehicle), because "bonnet" is most frequently linked to this page in Wikipedia.

In our setting, the average number of senses available in Wikipedia for each lexical unit, corresponding to linked pages, is 11.

6.2. Mapping analysis

We manually inspected a sample of 500 mappings between (l, F) pairs and Wikipedia pages. The pairs have been selected in order to maximize frame variability, so that every pair contains a different frame. Table 5 shows the results of the analysis.

Table 4
Output of Wikipedia mapping.

N. of (l, F) pairs	4,154
N. of mapped lexical units	3,800
N. of frames in the mapping	492
Different mappings w.r.t. the baseline	277

Table 5
Mapping analysis (sample of 500 mappings).

N. of exact matches	309
N. of exact matches in the *baseline*	280
Wikipage more *generic* than l	10
Wikipage more *specific* than l	35
N. of wrong matches	71
N. of wrong matches & missing sense	75

We compare this result with the most-frequent baseline, as defined in Section 6.1. Wrong mappings are cases in which the WSD system assigns to (l, F) the wrong sense, even if the right one is among the possible senses listed in Wikipedia. Also, we report in a separate row the number of wrong assignments due to the fact that the correct sense is missing in Wikipedia. This measure can be used to estimate the upper-bound accuracy of the mapping. Our analysis also includes the mappings in which (l, F) and W are semantically related, but the Wikipedia concept is either more general or more specific than (l, F). For example, (*belief.n*, Awareness) is mapped to the `Belief` page, which however describes the "belief" concept in a broader sense and includes the meaning of another (l, F) pair, namely (*belief.n*, Opinion). In few other cases, the relation between (l, F) and W is reversed. For example, (*enlargement.n*, Expansion) is mapped to `Enlargement_of_the_European_Union`.

6.3. Discussion

Table 5 shows that about 15% of the evaluated lexical units cannot be mapped to any page because the corresponding concept is not present in Wikipedia. This confirms our hypothesis that FrameNet and Wikipedia are linkable resources as far as nominal lexical units are concerned. Overall, our mapping covers 37% of FrameNet lexical units and 68% of all frames (Table 4). As we will show in Section 8, the latter result is more relevant than the coverage achieved over single lexical units, because the methodology to acquire new FrameNets in multilingual settings is performed on a frame basis. If at least one Wikipedia page is associated with a frame through a lexical unit, then all lexical units linked to such page can be acquired, regardless of their part of speech.

In general terms, inherent differences between FrameNet and Wikipedia affect the mapping of the two resources. The most evident distinction is the large number of concepts encoded in Wikipedia (3.82 million concepts) compared to the smaller set of lexical units (10,195 in FrameNet 1.3). Another relevant difference is that FrameNet is built upon semantic classes of concepts, the *frames*, while Wikipedia pages are less structured. Although the set of words that point to the same Wikipedia page can be seen as semantically related words, there is no design *a priori* of the characteristics of these classes, which are implicitly defined by different editors through the links. For example, professionals are often linked to the corresponding profession or activity in Wikipedia, while two distinct frames were created for Medical_specialties and Medical_professions in FrameNet. However, also in FrameNet this distinction is not consistently performed, since some other frames contain LUs denoting both persons and activities (see for example *assassin.n*, *assassination.n* and *assassinate.v* in Killing). Another main difference is the way the transitive and intransitive use of a verb is treated: in FrameNet it is usually modeled through frame alternation (e.g. Cause_to_make_noise and Make_noise), while this distinction is generally not made by Wikipedia editors.

With regard to our disambiguation strategy, other approaches could be investigated without relying on pseudo-contexts. For instance, lexical units can be disambiguated using the context of the FrameNet sentences in which they occur, that are similar to the contexts used for training. However, this approach would only account for the lexical units instantiated by at least one example sentence in FrameNet. A comparison between the two strategies shows that the approach based on pseudo-contexts can disambiguate 3800 lexical units, while the *example-based* approach covers only 2505 lexical units. A comparison showed that only 10% of the assignments made by both approaches do not match and neither of the two achieved significant improvement over the other in terms of precision. Therefore, we take into account only the mapping obtained using pseudo-contexts.

The example-based approach could benefit from the availability of a large number of sentences for each lexical unit. In order to acquire additional examples, each lexical unit in context could be replaced by the lexical units belonging to the same frame. However, this would partly lead to the creation of ungrammatical sentences, given that LUs in the same frame can have different parts of speech. Also antonymous lexical units may affect the quality of sentences created via substitution.

Table 6
Frame identification over seen lexical units. Symbols † and ‡ indicate significant differences relative to SEMAFOR ($p < 0.01$ and $p < 0.05$, respectively). Significance tests were computed using approximate randomization procedure.

	Precision	Recall	F_1
SEMAFOR	0.76	0.73	0.75
WIKI-BASED	$0.81^{‡}$	$0.25^{†}$	$0.38^{†}$
WIKI-BASED+WLM	0.74	$0.27^{†}$	$0.39^{†}$

7. English frame identification

The WSD system and the mapping between (l, F) pairs and Wikipedia pages can be straightforwardly used for automatic frame identification. Given an English sentence containing a lexical unit, we first disambiguate it by assigning a Wikipedia page, and then exploit the mapping to find the most appropriate frame.

The potential of our approach is greater for languages for which FrameNet is not available, because it performs frame identification using only Wikipedia for training. For English, supervised systems such as Shalmaneser [7] and SEMAFOR [8], which are trained directly on FrameNet, have already achieved good performances. However, we first test our approach on English because (i) we want to assess the difference between our methodology and standard supervised approaches, which are available only for English, and (ii) we want to assess if the performance of our approach decreases when moving from English to other languages.

7.1. Frame identification over seen lexical units

As a first step, we compare our methodology to the performance of SEMAFOR-1 [8], a state-of-the art frame semantic parser which applies a conditional log-linear model to prospected lexical units for frame identification. We choose this system because it significantly outperforms the other systems participating in the SemEval 2007 task for frame semantic parsing [8] and because it is freely available.[16]

7.1.1. Experimental setup

In order to compare the two approaches, we use the test set of SemEval 2010 task “Linking events and their participants in discourse” [60]. Although the task concerned argument labeling of given lexical units, we use this test set as a gold standard for frame identification. In particular, we take into account the subset of LUs listed in FrameNet 1.3, which is the training set for SEMAFOR-1 and the source data set for our mapping. This subset of the SemEval test set contains 1432 lexical units, divided into nouns (38%), verbs (51%), adjectives (9%) and other PoS including prepositions, adverbs and determiners (2%).

Note that the selection strategy adopted for lexical unit identification is different: in SEMAFOR a set of rules is applied to filter out bad candidates, for example some support predicates and prepositions. In our approach each term or multiword which was disambiguated is potentially seen as a LU candidate. However, we only include in the gold standard the lexical units that appear in FrameNet 1.3.

We test our methodology based on the mapping process described in Section 6. Then, we compare it with a second setting, where we exploit a Wikipedia-based similarity library to improve coverage. The basic intuition is that, even if we link a lexical unit l_i to a Wikipedia page w_i which is not included in the original mapping M, we can retrieve from M the page w_n that is most similar to w_i and assign to l_i the frame f_n that was originally mapped to w_n. To this end, we compute a similarity score between w_i and each $w \in M$ and select as the best match the page $w_n \in M$ with the highest similarity score. We apply the Wikipedia link-based measure (WLM) [61] available through the Wikipedia Miner Toolkit.[17] Given two pages, this relatedness measure takes into account the incoming and outgoing links of each page and assumes that two pages sharing many links are more similar than those containing different links. The measure of similarity between w_i and w_n ranges between 0 and 1, with the majority of cases being comprised between 0 and 0.5. The evaluation is performed using different cutoff values, which produce slightly different, but not significant, results. The values reported in Table 6 are obtained with a threshold set to zero.

7.1.2. Evaluation

As shown in Table 6, our methodology is affected by recall problems, since only lexical units linked to a Wikipedia page that appears in the mapping can be identified. In addition, our approach does not evenly cover all parts of speech (only 3 verbal LUs are correctly annotated). In English documents, the Wikipedia-based approach to frame identification cannot compete with supervised systems in the classification of seen lexical units. A frame identification model trained directly on the FrameNet database covers a smaller set of possible senses, which guarantees a better performance than a disambiguation

[16] http://www.ark.cs.cmu.edu/SEMAFOR/.
[17] http://wikipedia-miner.sourceforge.net/index.htm.

Table 7
Evaluation of frame annotation of unseen LUs. Symbols † and ‡ indicate significant differences relative to SEMAFOR ($p < 0.01$ and $p < 0.05$, respectively). Significance tests were computed using approximate randomization procedure.

	Precision	Recall	F_1
All frames and all LUs (AF–AL)			
SEMAFOR	0.28	0.14	**0.19**
WIKI-BASED	**0.34**†	0.09†	0.14†
WIKI-BASED+WLM	0.20†	**0.15**	0.17
All frames and nominal LUs (AF–NL)			
SEMAFOR	**0.52**	0.21	0.30
WIKI-BASED	0.38†	0.17‡	0.24†
WIKI-BASED+WLM	0.38†	**0.35**†	**0.36**‡
Restricted frame set and all LUs (RF–AL)			
SEMAFOR	0.25	0.12	0.16
WIKI-BASED	**0.39**†	0.13	0.19‡
WIKI-BASED+WLM	0.27	**0.21**†	**0.23**†
Restricted frame set and nominal LUs (RF–NL)			
SEMAFOR	**0.49**	0.18	0.26
WIKI-BASED	0.42‡	0.18	0.25
WIKI-BASED+WLM	0.40†	**0.37**†	**0.38**†

model trained on Wikipedia. Nevertheless, this analysis suggests that, when applying the Wikipedia-based methodology to other languages, good annotation precision can be expected, even if no direct supervision is possible.

7.2. Frame identification over unseen lexical units

The annotation of *unseen* lexical units is still an open problem, which is generally not tackled by current frame semantic parsers [6,7] or partially solved by relying on WordNet [10,26] and distributional models [13].

Our approach to frame identification does not require specific strategies for unseen lexical units, because the underlying disambiguation model remains the same as for seen lexical units. Compared to other approaches, no additional resources or large corpora are required, because any word or expression occurring in the training set of the WSD system is treated as a potential lexical unit and does not need to be included in FrameNet.

We evaluate our approach on this specific task by comparing it to SEMAFOR, which relies on lexico-semantic features partly extracted from WordNet to relate unseen with seen lexical units [62]. We choose SEMAFOR because it is the only system available that handles both seen and unseen lexical units, similar to our approach.

7.2.1. Experimental setup

Approaches dealing with the annotation of unseen lexical units were evaluated in a leave-one-out fashion using FrameNet 1.3 as gold standard [13,10]. We introduce a more realistic setting by classifying lexical units that are completely missing in FrameNet 1.3.

We extract all sentences that have been added to the FrameNet 1.5. database with respect to the previous release (version 1.3). The sentences introducing new LUs that belong to frames already present in FrameNet 1.3 are included in our test set. We discard lexical units belonging to frames that were not present in version 1.3, since the task of frame discovery is beyond the scope of this article.

The test set includes 2736 sentences containing 179 (l, F) pairs. In order to evaluate the impact of the mapping coverage on our approach, we also consider a smaller test set, including only frames that are represented in the mapping with Wikipedia pages (1976 sentences, 132 (l, F) pairs). We also evaluate nominal lexical units separately, both in the setting including all frames (880 sentences, 57 (l, F) pairs) and in the setting including only the frames in the mapping (828 sentences, 56 (l, F) pairs).

7.2.2. Evaluation

In Table 7, we report the evaluation of the frame assignment task over unseen lexical units on the whole test set (all frames and all LUs, *AF–AL*). We also take into account the three subsets mentioned in Section 7.2.1: the set comprising all frames and only nominal LUs (*AF–NL*), the set including the restricted set of frames occurring in the mapping and all LUs (*RF–AL*), and the third including only the restricted frame set and the nominal LUs (*RF–NL*).

The evaluation on the whole test set confirms the findings shown in Table 6: our system in the basic setting outperforms SEMAFOR with regard to classification precision, while recall is a major issue. The fact that some frames are not included in our mapping has a relevant impact on the results, since WIKI-BASED outperforms SEMAFOR on the *RF–AL* test set. In general, the small amount of mappings acquired between Wikipedia pages and (l, F) pairs is the main reason for the low recall. 90% of the missing assignments on the *AF–AL* test set are due to low mapping coverage, while only 10% were not disambiguated by the WSD system. The strategy to improve recall using WLM proves to be effective, since it achieves the best recall on every test set. With nominal lexical units, this implies a limited drop in precision (*RF–NL*), or no reduction at all (*AF–NL*) compared to WIKI-BASED. However, the drop is more dramatic if the similarity measure is applied to all lexical units.

7.3. Discussion

A major shortcoming of our system is that it recognizes and labels mainly nominal LUs. This depends on the set of (l, F) pairs that were originally mapped to Wikipedia, since it contains only nominal LUs. Besides, Wikipedia structure is based on concepts, which are typically expressed by nominal expressions. However, the analysis of the correct assignments over the whole test set in Table 7 (*AF–AL*) shows that 58% of the correct assignments concern nominal LUs, while 26% are verbs and 16% are adjectives. In other words, even if preference is given to nominal lexical units, other parts of speech are handled. The results obtained with SEMAFOR suggest that also this system suffers from lower performance when handling unseen verbal and adjectival LUs, since it shows a remarkable drop in precision and recall in the all-LUs setting.

In order to generalize from seen to unseen lexical units, two main approaches have been presented in past works: one relies on WordNet [10,26,13] and the other on distributional similarity [13]. A shortcoming of the first approach is that it can be applied only to unseen lexical units having a WordNet entry and is limited to languages for which WordNet is available. Burchardt et al. [10] evaluate this methodology with a leave-one-out strategy using FrameNet 1.2 and report 87% coverage and 39% precision (recall computed the same way as in our experiments is 34%). However, the authors define precision as a 'weak' measure of overlap with the gold standard, since it corresponds to the number of times the list of suggested frames for a given LU contains the gold standard frame. These measures would likely decrease in an evaluation setting based on standard precision and recall. Furthermore, the WordNet-based detour was evaluated on FrameNet 1.2, containing 600 frames, while FrameNet 1.3 used in our experiments contains 725 frames, which makes the assignment task more difficult.

Distributional approaches, on the other hand, require that large corpora are available for representing existing frames and unknown LUs in a semantic space. Using the BNC as reference corpus combined with WordNet, Pennacchiotti et al. [13] show that new lexical units can be added to existing frames achieving 0.43 accuracy and 0.95 coverage in a best-first model. However, this approach is not comparable to ours for several reasons: first, the experimental setting is different, because a lexical unit had to be re-assigned to a frame after it had been removed from all its original frames, as in [10]. The assignment was considered correct if *at least one* of the original frames was matched and was performed on a *per lemma* basis. This means that LU ambiguity, i.e. the possibility that a lemma corresponds to two LUs in two different frames, was not accounted for in the evaluation. As a result, the methodology in [13] cannot be applied to perform frame identification over running text, because it does not account for the different senses (equal to frames) that a LU can bear depending on the context in which it occurs. Furthermore, evaluation was performed on a reduced test set, including only frames with more than 2 LUs and candidate LUs with a frequency higher than 5 in the BNC (67% of FrameNet LUs).

As a preliminary conclusion, the experiments on English show that our approach can rival SEMAFOR when dealing with unseen LUs. However, the possibility to train a WSD model directly on FrameNet data guarantees a better performance when classifying seen LUs. Also, our methodology does not evenly cover all parts of speech, while SEMAFOR is better balanced in this respect. Our approach shows that neither WordNet nor large reference corpora are necessary for frame identification, because Wikipedia sense repository represents a valuable alternative to these resources. Therefore, the greatest potential of our methodology is expressed in multilingual settings, where language-specific versions of WordNet and FrameNet are not always available.

8. Creation of multilingual FrameNets

Many research projects have been recently devoted to the development of FrameNet for languages other than English. However, the collection of a large amount of hand-crafted annotation is an expensive and time-consuming process that slows down the growth of these initiatives. German FrameNet[18] is the only FrameNet-like resource that has been released with a coverage comparable to the English version. To overcome this bottleneck, (semi)- automatic techniques have been developed to collect FrameNet-like data in other languages by exploiting external resources such as WordNet [11,63], parallel corpora [64,34,65,66], and bilingual dictionaries [32]. However, the non-availability of such resources or their limited development for many languages became a strong motivation for us to investigate the use of Wikipedia.

Wikipedia includes (i) multilingual information aligned at concept level, similar to bilingual dictionaries, (ii) textual data in 282 languages, organized as in a comparable corpus, with the addition of links between documents, and (iii) an internal structure that allows the computation of similarity measures between concepts, similar to WordNet. Even if the latter is organized in a more rigorous and coherent way, Wikipedia combines the three aspects mentioned above, so that it is particularly suitable for NLP applications in multilingual settings, when other resources are scarcely available.

For the development of new FrameNets, we first use Wikipedia to acquire a set of pre-classified sentences and LUs and, second, to perform frame identification on multilingual documents. The basic assumption in our approach is that frames capture language-independent situations or events, and that the frame repository of English FrameNet can represent the

[18] http://www.coli.uni-saarland.de/projects/salsa/corpus/.

Table 8
Statistics on data extracted from Italian Wikipedia.

English Wikipedia articles in the mapping	3,818
Linked articles in Italian Wikipedia	1,866
N. of extracted sentences	610,397
Frames with at least 1 sentence	404
LU candidates	14,415

Table 9
Sentence and LU evaluation in Italian.

Accuracy (sentence extraction)	0.69
Accuracy (LU induction)	0.62
N. of new LU candidates	635
N. of correct candidates	396

backbone of FrameNets in other languages without major changes. This assumption has already been confirmed in different projects aimed at automatically acquiring data in other languages [20,32,67].

8.1. Acquisition of lexical units and example sentences

The goal of this phase is the automatic acquisition of LUs and example sentences in a non-English language L_n. For each English Wikipedia article W mapped to an (l, F) pair (Section 6), we retrieve the Wikipedia article W_n in L_n by following the cross-language link between W and W_n. In L_n, W_n is supposed to be either a translation of W or an alternative description of the same concept. If W_n exists, we extract all sentences S_n that contain a link to W_n from Wikipedia in L_n. We assume that S_n are example sentences of F in L_n. In addition, all terms linked to W_n are acquired as LUs. The same approach can be exploited in principle for all languages represented in Wikipedia.

8.1.1. Experimental setup

Experiments have been performed on the Italian Wikipedia dump of June 2010.[19] In order to increase the mapping coverage, we disambiguate nominal LUs in FrameNet using pseudo-contexts (See Section 6.1) in case they are not instantiated by an example sentence, otherwise, we apply the example-based disambiguation (see Section 6.3). The final mapping contains 3818 links, i.e. 18 more than the mapping based only on definitions.

We filter out all sentences shorter than 40 characters and the ones extracted for the PEOPLE_BY_ORIGIN frame. The first filter is applied to discard examples that are likely to be incomplete, as we need to create a first set of examples that possibly contain enough textual material for future annotation with semantic roles. The second filter has been introduced to tackle a problem that involves the links to nationalities and nations in Wikipedia. The mapping algorithm usually connects a LU belonging to the PEOPLE_BY_ORIGIN frame, which contains a list of nationalities, to the Wikipedia article about the corresponding nation. For example, the *Brit.n* LU is mapped to the `United_Kingdom` article, which is in turn linked to the Italian `Regno_Unito` article. Even though this can be considered conceptually correct, it represents a relevant problem for our task, because nationalities and nations belong to different frames. Many sentences are extracted from Wikipedia articles describing nations, because these are among the most linked articles. Therefore, the rule was necessary to reduce the negative impact of this annotation practice on the extraction task. Table 8 shows the statistics for the extracted data.

8.1.2. Evaluation

We manually evaluate 2000 examples consisting of 5 sentences randomly extracted from 400 linked Wikipedia articles, which in turn are randomly drawn. The evaluation is focused on the performance of the sentence extraction algorithm and aims at assessing if the given sentences could be included in the Italian version of F. We also check if the candidate LUs occurring in the evaluated sentences can be appropriately included in F. The results are shown in Table 9.

We computed Cohen's kappa statistic [68] over a random sample of 200 sentences. Agreement between two judges achieves 0.82, which as a rule of thumb is seen as a very good agreement and represents a solid basis for our evaluation.

A manual inspection of the evaluation set shows that 34% of the wrong sentences were extracted due to disambiguation errors. Around 44% of the mistakes depend on how frames are structured compared to Wikipedia pages: frames are often built around specific parts of speech, while this distinction is not found in Wikipedia. For instance TOXIC_SUBSTANCE lists harmful substances such as *poison.n* and *venom.n*, while CAUSE_HARM groups verbs describing harmful actions such as *to poison.v*. In Wikipedia, these terms are all linked to the `Poison` page, regardless of their part of speech. In the remaining cases (around 22%), mistakes depend on cross- and intra-language links in Wikipedia. Specifically, several fine-grained sense distinctions in English are missing in Italian, since the extension of Italian Wikipedia is around 23% of the English

[19] http://download.wikimedia.org/itwiki/20100624/.

Table 10
N. of correct LUs acquired for each frame, grouped by part of speech.

ATTEMPT_SUASION	2 n
CHANGE_POS._ON_A_SCALE	4 n, 1 v, 3 a
DEATH	17 n, 8 v, 6 a
DEPARTING	6 n, 3 v, 2 a
SELF_MOTION	9 n, 3 v, 2 a

Table 11
Frame annotated data in MultiSemCor.

	WIKI-BASED	WN-BASED
N. of annotated LU occurrences	17,714	23,872
N. of acquired Italian (l, F) pairs	1,708	3,380
with a *nominal* LU	1,097	1,341
with a *verbal* LU	272	1,525
with an *adjectival* LU	307	512
with an *adverbial* LU	32	2
Frames with at least 1 acquired LU	269	530

one. Wrong sentences have also been extracted because of wrong internal links. For example, sentences dealing with the newspaper *La Nazione* have been linked to the `Nazione` (*Nation*) page.

As shown in Section 7, our approach is noun-centered, because it is constrained by the design of Wikipedia based on concepts. This holds also for LU induction in the multilingual setting, because most of acquired LUs are nouns. Nevertheless, the 396 LUs evaluated as correct in our data set include also 2 adverbs, 15 verbs and 30 adjectives. In order to further investigate the impact of our noun-centered approach on LU acquisition, we randomly select 5 frames that in FrameNet are mostly characterized by verbal lexical units (the proportion of verbal lexical units w.r.t. nominal ones being > 2). We extract from the data set reported in Table 8 the sentences acquired for these frames. Then, we analyze the part of speech of the correct LUs acquired from these sentences through the induction process.

The statistics reported in Table 10 confirm that nominal LUs are predominant in the acquisition phase, but also that our methodology can induce other LU types, except for ATTEMPT_SUASION. This is in line with the analysis of the system behavior reported for the annotation of unseen LUs in English (Section 7.2).

8.2. Development of multilingual frame assignment systems

The backbone of our methodology can be also exploited to perform multilingual frame identification. The components needed are the same as for English: a supervised WSD system, Wikipedia, and the Wikipedia–FrameNet mapping. We show that their combination can potentially lead to the development of a frame annotation system for any language available in Wikipedia.

The methodology is similar to the one applied to English (Fig. 3): we first train the WSD system using Wikipedia in the language L_n, following the procedure described in Section 5. Second, we disambiguate the terms in a document d in the language L_n. For each Wikipedia page W_n that is linked to a term t_n in d, we retrieve the English article W_e using the cross-language link between W_n and W_e. Finally, we retrieve from the Wikipedia–FrameNet mapping the frame f_e mapped to article W_e, and assign it to term t_n.

8.2.1. Experimental setup

We test the methodology on the Italian part of the MultiSemCor corpus [69]. This data set contains 12,843 sentences enriched with synset labels, that were automatically transferred from the English counterpart with a precision of 0.86. Tonelli and Pighin [11] employ the same corpus to test a procedure that maps a WordNet synset to the corresponding frame, based on a set of lexico-semantic features. Since it is the only attempt to acquire and evaluate lexical units and example sentences on a real corpus of Italian, we evaluate our frame assignment approach on the same data. However, in [11] gold English synsets were given and the Italian ones had been automatically acquired via word alignment. On the contrary, our methodology includes disambiguation. Therefore, the classification accuracy reported by [11] represents a sort of upper bound for our approach.

Statistics about the Italian data annotated with our methodology are reported in Table 11 (WIKI-BASED). We report also the same statistics related to the data extracted by [11] (WN-BASED).

Most of the acquired LUs are nouns, but other PoS are also included. The Wikipedia-based approach covers a smaller range of LUs and frames, but more examples are acquired on average for each LU than through WordNet, i.e. 10 vs. 7 sentences for each (l, F) pair.

Table 12
Evaluation of frame assignment on MultiSemCor.

Nominal LUs in the evaluation set	674
correctly classified	492
Verbal LUs in the evaluation set	217
correctly classified	47
Adjectival LUs in the evaluation set	91
correctly classified	36
Adverbial LUs in the evaluation set	18
correctly classified	5
Classification accuracy	**0.58**

Table 13
Comparison between 10 most populated WordNets and Wikipedia.

Language	N. of WordNet synsets	Project name	N. of Wikipedia articles
English	155,287	Princeton WN	3,817,361
Polish	73,839	plWordNet	848,759
German	69,594	GermaNet	1,324,259
Spanish	57,424	MultiWordNet	849,263
Japanese	57,238	Japanese WN	781,433
Dutch	44,015	EuroWordNet	908,386
Italian	38,658	MultiWordNet	867,277
French	32,351	WOLF	1,181,260
Hindi	26,208	HindiWordNet	100,549
Romanian	20,191	MultiWordNet	169,472

8.2.2. Evaluation

In order to compare our approach to [11], our evaluation should be applied to the same data set. However, in [11] only 200 instances were manually evaluated. After running our system on the same data set, we noticed that only 42 instances were annotated with both approaches. This number is too small to allow for an accurate evaluation and comparison of the two systems. Therefore, we decided to manually evaluate another part of the corpus, i.e. 1000 sentences with one annotated instance each. This data set is built in order to include some sentences from each of the documents in MultiSemCor, which cover several topics. For each labeled instance, an annotator is asked to assign a positive judgment if the frame label given by the system is correct, and a negative judgment otherwise. Since there are no reference frames in Italian, the annotator is allowed to look up frame descriptions and example sentences in the English FrameNet. More details on the frame assignment methodology are given in [70]. Evaluation results are reported in Table 12.

Classification accuracy reported by [11] is 0.70, but it is based on a smaller evaluation set. Given that the synsets are already defined, while our methodology includes disambiguation, we conclude that our approach is quite promising. In order to apply the methodology by [11] to any document, an unsupervised WSD system based on WordNet is needed. Given the performance of existing systems, this integration would lead to a significant drop in accuracy.[20] Also the coverage statistics reported in Table 11 for the WordNet-based approach would be considerably different in a real application scenario, since recall of the best unsupervised WSD systems is still below 60%.

Precision on (seen) LUs in English is 0.81 (Section 7). Although the results on two different languages are not directly comparable, the difference in accuracy reflects the impact of the errors introduced in the Italian-specific setting, i.e., the mistakes of the WSD system trained on Italian and the quality of the links between English and Italian Wikipages.

8.3. Discussion

Other presented approaches have been aimed at acquiring an initial set of pre-classified sentences and LUs in a new language. However, their shortcoming is that they are strongly language-dependent, because they require specific NLP components and resources. Both Tonelli and Pighin [11] and de Cao et al. [27] acquire Italian LUs based on MultiWordNet, but the same procedure can be applied only to a limited number of languages. EuroWordNet [31] includes more languages than MultiWordNet, but the English synsets are not strictly aligned across languages, which makes the WordNet-based methodology not applicable.

[20] Agirre et al. [71] report that, in the last SemEval-2010 task on WordNet-based unsupervised WSD in a specific domain, the best performing system for English achieved P 0.57 and R 0.55, while the top-ranked system for Italian scored P 0.53 and R 0.53. Although the accuracy of the best supervised systems is higher, being around 75% [57], they are not applicable in practical applications due to the high cost of creating and maintaining training data.

Wikipedia covers 282 languages,[21] while around 60 are available in WordNet.[22] This represents a clear advantage of our methodology compared to WordNet-based approaches. Also the coverage of the language-specific versions of Wikipedia clearly exceeds WordNet, as confirmed by the data in Table 13, comparing the 10 most populated WordNets with the corresponding Wikipedia version. We also report the name of each specific WordNet project, since different projects are developing WordNets in the same language (e.g., MultiWordNet and EuroWordNet for Italian and Spanish).

In absolute terms, the extraction of frame example sentences from Italian Wikipedia leads to the creation of a data set comprising 610,397 sentences with 14,415 candidate LUs. Through MultiSemCor (see Section 8.2.1) and the mapping between frames and MultiWordNet synsets, [11] extract 23,872 Italian sentences and 6429 candidate LUs. Note that 3049 LUs are acquired by [11] directly through the synsets, i.e., they are not instantiated by examples. This confirms that Wikipedia allows for the acquisition of remarkably more data than WordNet, supporting lexicographers and annotators in the development of new FrameNets.

As for frame identification on new languages, it is an almost unexplored task, given that existing frame semantic parsers require a large set of training data. Our methodology is the first attempt to handle the task without frame-specific supervision, so that it can be extended to all languages in Wikipedia. To our knowledge, the only available system handling non-English documents is Shalmaneser [7], which was trained using the German SALSA corpus [9]. The system labels mainly verbs, since in SALSA 97% of the LUs are verbs. Our methodology can complement German FrameNet, enriching the existing frames with more nominal lexical units and example sentences. It can also integrate Shalmaneser output by assigning a frame label to unseen LUs in German documents.

9. Conclusions

Our goal in this article has been to present an approach to frame identification and frame example acquisition for all languages in Wikipedia. In most of the cases, training data with high-quality frame annotation are not available. Therefore, we have developed a methodology avoiding direct supervision over FrameNet data. Research in this direction is crucial to alleviate performance degradation, when dealing with different domains and languages. Furthermore, in the FrameNet paradigm, frame annotation is of paramount importance, because it is a prerequisite for semantic role labeling, and its quality impacts on the whole annotation task.

Experiments on English showed that, compared to supervised approaches, our methodology achieves good precision but low recall (Section 7). A first attempt to alleviate this problem using a basic Wikipedia-based similarity showed a promising improvement. In the future, we plan to further improve recall by exploiting other Wikipedia-based similarity metrics [72].

The quality of the acquired data in Italian showed that our method is particularly promising in multilingual settings. It can be exploited by editors and lexicographers to develop new FrameNets in a semi-automatic way, especially when other lexical resources are not available. Yet, there is still room for improvement. In particular, recall can be boosted by increasing the number of inter-language links. Past works showed that it is possible to retrieve Wikipedia pages expressing the same concept in different languages, even if they are not explicitly connected by a link [73]. Experiments in this direction will increase the number of inter-language links through which the FrameNet–Wikipedia mapping is accessed.

Finally, research on lexical units would benefit from a deeper understanding of the semantics of their arguments, and vice versa. Therefore, we are developing a unified methodology in order to perform both frame identification and semantic role annotation using the same WSD system. In this framework, FrameNet, Wikipedia and WordNet are further integrated. A preliminary study on the semantics of role fillers based on the three resources confirmed that the approach is promising [74]. In the near future, we will investigate the best strategy to combine LU and semantic role information, and evaluate a frame semantic parser based on Wikipedia against state-of-the-art systems.

Acknowledgements

We would like to thank the three anonymous reviewers for their comments and suggestions. We are also thankful to Ryan Bohler for proofreading the final article version.

References

[1] C.F. Baker, C.J. Fillmore, J.B. Lowe, The Berkeley FrameNet project, in: Proceedings of the 17th International Conference on Computational Linguistics (COLING-98), Montreal, Quebec, Canada, 1998, pp. 86–90.

[2] C. Fillmore, C. Johnson, M.R.L. Petruck, Background to FrameNet, International Journal of Lexicography 16 (2003) 235–250.

[3] D. Shen, M. Lapata, Using semantic roles to improve question answering, in: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL-07), Prague, CZ, 2007, pp. 12–21.

[4] B. Rink, S. Harabagiu, UTD: Classifying semantic relations by combining lexical and semantic resources, in: Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-3), Uppsala, Sweden, 2010, pp. 252–255.

[5] R.B. Aharon, I. Szpektor, I. Dagan, Generating entailment rules from FrameNet, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10), Stroudsburg, PA, USA, 2010, pp. 241–246.

[21] http://meta.wikimedia.org/wiki/List_of_Wikipedias.

[22] http://www.globalwordnet.org/gwa/wordnet_table.html.

[6] B. Coppola, A. Moschitti, A general purpose FrameNet-based shallow semantic parser, in: Proceedings of the 7th Language Resources and Evaluation Conference (LREC-10), La Valletta, Malta, 2010, pp. 19–21.

[7] K. Erk, S. Padó, Shalmaneser – a toolchain for shallow semantic parsing, in: Proceedings of the 5th Language Resources and Evaluation Conference (LREC-06), Genoa, Italy, 2006, pp. 527–532.

[8] D. Das, N. Schneider, D. Chen, N.A. Smith, Probabilistic frame-semantic parsing, in: Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-10), Stroudsburg, PA, USA, 2010, pp. 948–956.

[9] A. Burchardt, A. Frank, S. Padó, M. Pinkal, The SALSA corpus: a German corpus resource for lexical semantics, in: Proceedings of the 5th Language Resources and Evaluation Conference (LREC-06), Genoa, Italy, 2006, pp. 969–974.

[10] A. Burchardt, K. Erk, A. Frank, A WordNet detour to FrameNet, in: B. Fisseni, H. Schmitz, B. Schröder, P. Wagner (Eds.), Sprachtechnologie, mobile Kommunikation und linguistische Resourcen, Peter Lang, Frankfurt am Main, Germany, 2005, pp. 408–421.

[11] S. Tonelli, D. Pighin, New features for FrameNet–WordNet mapping, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CONLL-09), Boulder, CO, USA, 2009, pp. 219–227.

[12] L. Shi, R. Mihalcea, Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing, in: Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-05), Springer, Mexico City, Mexico, 2005, pp. 100–111.

[13] M. Pennacchiotti, D. de Cao, R. Basili, D. Croce, M. Roth, Automatic induction of FrameNet lexical units, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-08), Waikiki, Honolulu, Hawaii, 2008, pp. 457–465.

[14] M. Pennacchiotti, D. de Cao, P. Marocco, R. Basili, Towards a vector space model for FrameNet-like resources, in: Proceedings of the 6th Language Resources and Evaluation Conference (LREC-08), Marrakech, Morocco, 2008, pp. 790–796.

[15] S. Tonelli, C. Giuliano, Wikipedia as frame information repository, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-09), Singapore, 2009, pp. 276–285.

[16] C.J. Fillmore, Frames and the semantics of understanding, Quaderni di Semantica IV (2) (1985) 222–254.

[17] C.J. Fillmore, Frame semantics and the nature of language, in: Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language, Blackwell Publishing, 1976, pp. 20–32.

[18] H.C. Boas, Semantic frames as interlingual representations for multilingual lexical databases, International Journal of Lexicography 18 (4) (2005) 445–478.

[19] A. Burchardt, M. Pennacchiotti, S. Thater, M. Pinkal, Assessing the impact of frame semantics on textual entailment, Natural Language Engineering 15 (4) (2009) 527–550 (Special Issue on Textual Entailment).

[20] C. Subirats, Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon, in: H.C. Boas (Ed.), Multilingual FrameNets in Computational Lexicography: Methods and Applications, in: Trends in Linguistics, Mouton de Gruyter, 2009, pp. 135–162.

[21] S. Tonelli, Semi-automatic techniques for extending the FrameNet lexical database to new languages, Ph.D. thesis, Dept. of Language Sciences, Università Ca' Foscari, Venezia, Italy, 2010.

[22] M.R. Petruck, Typological considerations in constructing a Hebrew FrameNet, in: H.C. Boas (Ed.), Multilingual FrameNets in Computational Lexicography: Methods and Applications, in: Trends in Linguistics, Mouton de Gruyter, 2009, pp. 183–205.

[23] K.H. Ohara, Lexicon, grammar, and multilinguality in the Japanese FrameNet, in: Proceedings of the 6th Language Resources and Evaluation Conference (LREC-08), Marrakech, Morocco, 2008, pp. 3264–3268.

[24] K.H. Ohara, Frame-based contrastive lexical semantics in Japanese FrameNet: the case of risk and kakeru, in: H.C. Boas (Ed.), Multilingual FrameNets in Computational Lexicography: Methods and Applications, in: Trends in Linguistics, Mouton de Gruyter, 2009, pp. 163–182.

[25] C. Fellbaum, WordNet. An Electronic Lexical Database, MIT Press, 1998.

[26] R. Johansson, P. Nugues, Using WordNet to extend FrameNet coverage, in: Proceedings of the Workshop on Building Frame-Semantic Resources for Scandinavian and Baltic Languages, at NODALIDA, Tartu, Estonia, 2007, pp. 27–30.

[27] D. de Cao, D. Croce, M. Pennacchiotti, R. Basili, Combining word sense and usage for modeling frame semantics, in: Proceedings of the Symposium on Semantics in Systems for Text Processing (STEP-08), Venice, Italy, 2008, pp. 85–101.

[28] R. Green, B.J. Dorr, P. Resnik, Inducing frame semantic verb classes from WordNet and LDOCE, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 2004, pp. 375–382.

[29] E. Pianta, L. Bentivogli, C. Girardi, MultiWordNet: Developing an aligned multilingual database, in: First International Conference on Global WordNet, Mysore, India, 2002, pp. 292–302.

[30] M. Crespo, P. Buitelaar, Domain-specific English-to-Spanish translation of FrameNet, in: Proceedings of the 6th Language Resources and Evaluation Conference (LREC-08), Marrakech, Morocco, 2008, pp. 1470–1473.

[31] P. Vossen (Ed.), EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Springer, 1998.

[32] C. Mouton, G. de Chalendar, B. Richert, FrameNet translation using bilingual dictionaries with evaluation on the English–French pair, in: Proceedings of the 7th Language Resources and Evaluation Conference (LREC-10), La Valletta, Malta, 2010, pp. 20–27.

[33] B. Chen, P. Fung, Automatic construction of an English–Chinese bilingual FrameNet, in: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics HLT/NAACL-04, Boston, USA, 2004, pp. 29–32.

[34] S. Padó, M. Lapata, Cross-lingual bootstrapping of semantic lexicons: The case of FrameNet, in: Proceedings of the 20th National Conference on Artificial, Intelligence (AAAI-05), vol. 3, AAAI Press, 2005, pp. 1087–1092.

[35] M. Ruiz-Casado, E. Alfonseca, P. Castells, Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets, Advances in Web Intelligence 3528 (2005) 380–386.

[36] O. Medelyan, C. Legg, Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense, in: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy, vol. 8, 2008, pp. 13–18.

[37] D. Lenat, CYC: A large-scale investment in knowledge infrastructure, Communications of the ACM 38 (11) (1995) 33–38.

[38] S. Sarjant, C. Legg, M. Robinson, O. Medelyan, "All you can eat" ontology-building: Feeding Wikipedia to Cyc, in: Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies (WI-IAT'09), 2009, pp. 341–348.

[39] F. Suchanek, G. Kasneci, G. Weikum, Yago: A core of semantic knowledge, in: Proceedings of the 16th International World Wide Web Conference, ACM, Banff, Alberta, Canada, 2007, pp. 697–706.

[40] R. Navigli, S. Ponzetto, BabelNet: Building a very large multilingual semantic network, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 216–225.

[41] A. Csomai, R. Mihalcea, Linking documents to encyclopedic knowledge, IEEE Intelligent Systems 23 (5) (2008) 34–41 (Special Issue on Natural Language Processing for the Web).

[42] D. Milne, I.H. Witten, Learning to link with Wikipedia, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM-08), ACM, New York, NY, USA, 2008, pp. 509–518.

[43] S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti, Collective annotation of Wikipedia entities in web text, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-09), ACM, New York, NY, USA, 2009, pp. 457–466.

[44] E. Agirre, P.G. Edmonds, Word Sense Disambiguation: Algorithms and Applications, Springer, 2006.

[45] R. Navigli, Word sense disambiguation: a survey, ACM Computing Surveys 41 (2) (2009) 1–69.

[46] C. Giuliano, A.M. Gliozzo, C. Strapparava, Kernel methods for minimally supervised WSD, Computational Linguistics 35 (4) (2009) 513–528.

[47] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[48] D. Yarowsky, Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French, in: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, 1994, pp. 88–95.

[49] H. Lodhi, J. Shawe-Taylor, N. Cristianini, C. Watkins, Text classification using string kernels, Journal of Machine Learning Research 2 (3) (2002) 419–444.

[50] C. Saunders, H. Tschach, J. Shawe-Taylor, Syllables and other string kernel extensions, in: Proceedings of 19th International Conference on Machine Learning (ICML02), 2002, pp. 530–537.

[51] N. Cancedda, E. Gaussier, C. Goutte, J. Renders, Word-sequence kernels, Journal of Machine Learning Research 32 (6) (2003) 1059–1082.

[52] N. Cristianini, H. Lodhi, J. Shawe-Taylor, Latent semantic kernels, Journal of Intelligent Information Systems 18 (2) (2002) 1–27.

[53] R. Mihalcea, Using Wikipedia for automatic word sense disambiguation, in: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL/HLT-07), Rochester, New York, USA, 2007, pp. 196–203.

[54] T. Zesch, C. Müller, I. Gurevych, Extracting lexical semantic knowledge from Wikipedia and Wiktionary, in: Proceedings of the 6th Language Resources and Evaluation Conference (LREC-08), Marrakech, Morocco, 2008, pp. 1646–1652.

[55] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[56] R. Mihalcea, P. Edmonds (Eds.), Proceedings of SENSEVAL-5: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Texts, Association for Computational Linguistics, Barcelona, Spain, 2004.

[57] S. Pradhan, E. Loper, D. Dligach, M. Palmer, SemEval-2007 Task-17: English lexical sample, SRL and all words, in: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 87–92.

[58] L. Bentivogli, P. Forner, C. Giuliano, A. Marchetti, E. Pianta, K. Tymoshenko, Extending English ACE 2005 corpus annotation with ground-truth links to Wikipedia, in: Proceedings of COLING 2010 Workshop "The People's Web Meets NLP: Collaboratively Constructed Semantic Resources", Beijing, China, 2010, pp. 19–26.

[59] P.N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia spotlight: Shedding light on the web of documents, in: Proceedings of the 7th International Conference on Semantic Systems (I-Semantics), Graz, Austria, 2011, pp. 1–8.

[60] J. Ruppenhofer, C. Sporleder, R. Morante, C.F. Baker, M. Palmer, SemEval-2010 Task 10: Linking events and their participants in discourse, in: Proceedings of the NAACL–HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, Boulder, CO, USA, 2009, pp. 106–111.

[61] I. Witten, D. Milne, An effective, low-cost measure of semantic relatedness obtained from Wikipedia links, in: Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA, 2008, pp. 25–30.

[62] D. Das, N. Schneider, D. Chen, N.A. Smith, SEMAFOR 1.0: A probabilistic frame-semantic parser, Tech. Rep. CMU-LTI-10-001, Carnegie Mellon University, 2010.

[63] P. Annesi, R. Basili, Cross-lingual alignment of FrameNet annotations through hidden Markov models, in: Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-10), Iasi, Romania, 2010, pp. 12–25.

[64] S. Padó, M. Lapata, Cross-lingual annotation projection of semantic roles, Journal of Artificial Intelligence Research 36 (2009) 307–340.

[65] S. Padó, G. Pitel, Annotation précise du français en sémantique de roles par projection cross-linguistique, in: Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN-07), Toulouse, France, 2007, pp. 271–280.

[66] S. Tonelli, E. Pianta, Frame information transfer from English to Italian, in: Proceedings of the 6th Language Resources and Evaluation Conference (LREC-08), Marrakech, Morocco, 2008, pp. 2252–2256.

[67] K.H. Ohara, S. Fujii, H. Saito, S. Ishizaki, T. Ohori, R. Suzuki, The Japanese FrameNet project: A preliminary report, in: Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING-03), Halifax, Canada, 2003, pp. 249–254.

[68] J. Carletta, Assessing agreement on classification tasks: The kappa statistic, Computational Linguistics 22 (2) (1996) 249–254.

[69] L. Bentivogli, E. Pianta, Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor corpus, Natural Language Engineering 11 (3) (2005) 247–261 (Special Issue on Parallel Texts).

[70] S. Tonelli, G. Riccardi, Guidelines for annotating the LUNA corpus with frame information, Tech. Rep. DISI-10-017, Department of Information Engineering and Computer Science, University of Trento, 2010.

[71] E. Agirre, O. López de Lacalle, C. Fellbaum, S.K. Hsieh, M. Tesconi, M. Monachini, et al., SemEval-2010 Task 17: All-words word sense disambiguation on a specific domain, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 75–80.

[72] M. Strube, S.P. Ponzetto, WikiRelate! Computing semantic relatedness using Wikipedia, in: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), Boston, USA, 2006, pp. 1419–1424.

[73] P. Sorg, P. Cimiano, Enriching the crosslingual structure of Wikipedia – a classification-based approach, in: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WikiAI'08), 2008, pp. 1–6.

[74] V. Bryl, S. Tonelli, C. Giuliano, L. Serafini, A novel FrameNet-based resource for the semantic web, in: Proceedings of the 27th ACM Symposium on Applied Computing, Association for Computing Machinery, Riva del Garda, Italy, 2012, pp. 360–365.