To Exhibit is not to Loiter: A Multilingual, Sense-Disambiguated Wiktionary for Measuring Verb Similarity

Christian M. Meyer[1] *Iryna Gurevych*[1,2]
(1) Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
(2) Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
http://www.ukp.tu-darmstadt.de

ABSTRACT
We construct a new multilingual lexical resource from Wiktionary by disambiguating semantic relations and translations. For this task, we propose and evaluate an automatic disambiguation method that outperforms previous approaches significantly. We additionally introduce a method for inferring new semantic relations based on the disambiguated translations. Our resource fills the gap between expert-built resources suffering from high cost and small size and Wikipedia-based resources that are restricted to encyclopedic knowledge about nouns. We demonstrate this by applying our new resource to measuring monolingual and cross-lingual verb similarity. For the latter, our resource yields better results than Wikipedia and expert-built multilingual wordnets. We make our final resource and the evaluation datasets publicly available.

TITLE AND ABSTRACT IN GERMAN

Ein mehrsprachiges, lesartendisambiguiertes Wiktionary zur Bestimmung von Verbähnlichkeiten

Der vorliegende Beitrag beschreibt die Gewinnung einer neuen, mehrsprachigen lexikalischen Ressource aus Wiktionary-Daten, die durch Disambiguierung von semantischen Relationen und Übersetzungen entsteht. Zu diesem Zweck definieren und evaluieren wir eine automatische Methode zur Lesartendisambiguierung, die frühere Ansätze signifikant übertrifft. Wir stellen ferner eine Methode vor, um neue semantische Relationen auf Basis der disambiguierten Übersetzungen zu inferieren. Unsere Ressource schließt die Lücke zwischen von Experten erstellten Wissensquellen, die unter ihrer oft geringen Größe aber hohen Erstellungskosten leiden, und Wikipedia-basierten Ressourcen, die nahezu ausschließlich enzyklopädisches Wissen zu Substantiven enthalten. Beim Einsatz unserer neuen Ressource zur Bestimmung von einsprachigen und zweisprachigen Verbähnlichkeiten erreichen wir im letzteren Fall bessere Ergebnisse als für Wikipedia und die Expertenressourcen. Wir veröffentlichen unsere Ressource und die Evaluierungsdatensätze für zukünftige Forschungsarbeiten.

KEYWORDS: Wiktionary, Lexical Resource, Semantic Relation, Translation, Word Sense Disambiguation, Verb Similarity.

KEYWORDS IN GERMAN: Wiktionary, Lexikalische Ressource, Semantische Relation, Übersetzung, Lesartendisambiguierung, Verbähnlichkeit.

Proceedings of COLING 2012: Technical Papers, pages 1763–1780,
COLING 2012, Mumbai, December 2012.

1 Introduction

Motivation. The advancing globalization and the permeation of the internet in our daily lives raises a strong demand for multilingual applications, such as machine translation, cross-lingual question answering, or information retrieval. Traditional multilingual approaches are knowledge-based using bilingual dictionaries (Neff and McCord, 1990) or multilingual wordnets (Tufiş et al., 2004). To date, these approaches are getting more and more replaced by statistical translation models, although it has been found that multilingual resources have the ability to substantially contribute to the performance of a system (Oepen et al., 2007; Herbert et al., 2011). One reason for the knowledge-based approaches being rarely employed is the challenging construction process of multilingual resources. They are either manually compiled by professional translators or lexicographers or automatically generated from large amounts of unstructured data. The former usually results in small resources due to the time and cost intensive work, whereas the latter often reaches only a limited quality. Although Wikipedia has been found as a promising alternative for obtaining multilingual knowledge (Medelyan et al., 2009), it is almost entirely restricted to nouns and focuses on encyclopedic rather than lexical-semantic knowledge.

Contribution. In this paper, we will explore the collaborative online lexicon Wiktionary[1] and how it can be used as a multilingual resource. Similar to Wikipedia, the contents in Wiktionary are edited by a large community of Web users. This collaborative construction approach, known as the "Wisdom of Crowds", yields very large resources. At the same time, this assures a considerable quality, as the numerous authors can quickly revise erroneous or unclear entries. Wiktionary offers a broad range of lexical-semantic knowledge including sense definitions, semantic relations, and translations. It fills the gap between the small, expert-built wordnets and the large Wikipedia-based resources restricted to nouns.

The contribution of our paper is threefold: (i) We propose and evaluate a method for disambiguating semantic relations and translations in Wiktionary; (ii) we infer new semantic relations based on the disambiguation result and create a novel sense-disambiguated Wiktionary that we make freely available; (iii) we demonstrate the usefulness of our new sense-disambiguated resource by employing it for calculating cross-lingual verb similarity. Measuring verb similarity is often a crucial technique for information extraction or (cross-lingual) question answering systems. In this paper, we experiment with English and German even though our methods can generally be adapted to over 170 languages covered by Wiktionary.

Overview. The English Wiktionary consists of about 475,000, the German Wiktionary of about 73,000 *word senses*.[2] For each of these word senses, multiple *semantic relations* (i.e., synonymy, antonymy, hyponymy, etc.) and *translations* may be encoded. We will use *relation* henceforth to refer to both semantic relations and translations and use the terms *source* and *target* to denote the endpoints of a relation. The Wiktionary entry for *(to) hang* distinguishes, for instance, fifteen word senses. The eighth word sense is defined as *"to exhibit (an object)"* with synonymy relations targeting at *exhibit* and *show* and translations into German *ausstellen*, French *exposer*, Dutch *ophangen*, and other languages.

The target of a relation is encoded using word forms. Thus, it remains underspecified which word sense a relation is pointing to. The synonym *exhibit* of the eighth word sense of *hang* can, for example, refer to the meaning of displaying something (e.g., exhibiting a drawing) or

[1] `http://www.wiktionary.org`

[2] All statistics are based on Wiktionary data of April 2011 accessed using JWKTL (Zesch et al., 2008a).

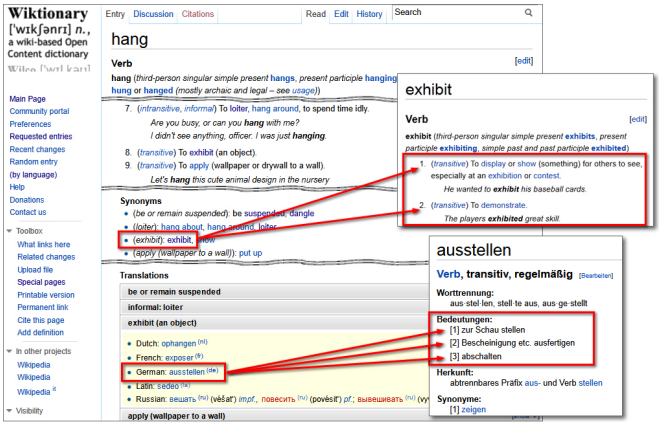

Figure 1: The synonym *(to) exhibit* of the English Wiktionary entry *(to) hang* and its German translation *ausstellen* have multiple possible target word senses.

demonstrating a skill (e.g., exhibiting a talent in acting). For humans, it is easy to recognize that *hang* is synonymous to the first word sense of *exhibit*, but not to the second. Natural language processing applications, however, cannot disambiguate such relations easily. The same applies to translations: The German *ausstellen* has, for instance, a meaning of (1) exhibiting an object, (2) certificating a document and (3) turning off smth. Figure 1 illustrates this kind of underspecification. In Section 3, we propose a solution to this issue by automatically disambiguating the semantic relations and translations in Wiktionary. Sense-disambiguated relations are a necessary precondition for many applications, such as computing semantic relatedness by measuring path lengths (Budanitsky and Hirst, 2006): if undisambiguated relations were used, then *exhibit* and *loiter* would be highly related as they both have a relation to *hang*.

Besides the information inherently found in Wiktionary, we infer new semantic relations based on our disambiguated translations. This is particularly useful for the English Wiktionary, which encodes only about 26,000 semantic relations (compared to 290,000 in the German edition). With our inference method, we are able to increase the number of semantic relations for the English language by almost ten times. Section 4 describes our inference method and provides statistics of our new resource. In Section 5, we apply this new resource to calculating monolingual and cross-lingual verb similarity as one example use case in the scope of our work. Thereby, we show that our resource is comparable to expert-built resources in the monolingual experiments and that it outperforms them in the cross-lingual setting by a large margin.

2 Related Work

The most closely related areas of work are the construction of multilingual resources, the disambiguation of relations and the inference of new semantic relations.

Multilingual resource construction. The most prominent multilingual resources are EuroWordNet (Vossen, 1998), BalkaNet (Stamou et al., 2002), and MultiWordNet (Pianta et al.,

2002). All of them are professionally crafted and provide a well-structured network of word senses and relations. Despite their high quality, the sizes vary largely. For English and German, there are, for instance, only 16,347 shared word senses in EuroWordNet (the only one encoding this language pair).[3] Another drawback of these wordnets is their high development cost which hinders the large-scale manual extension of their contents. Wikipedia-based multilingual resources is another strand of research. Well-known works are Yago (Suchanek et al., 2007), DBpedia (Bizer et al., 2009), and WikiNet (Nastase et al., 2010), which mostly differ in their structure and the way they extract the data. The bulk of knowledge in Wikipedia is, however, of encyclopedic nature, whereas our work aims at lexical-semantic knowledge.

The two most closely related research efforts to ours are Universal WordNet (de Melo and Weikum, 2009) and BabelNet (Navigli and Ponzetto, 2010). The former uses WordNet for bootstrapping a multilingual resource based on combined evidence found in existing wordnets, parallel corpora, and machine-readable dictionaries. It incorporates (undisambiguated) Wiktionary translations, but solely relies on semantic relations taken from WordNet. BabelNet aligns WordNet and Wikipedia at the level of word senses. Although this yields a large resource, the additional information from Wikipedia is almost entirely about nouns – there are hence no translations for verbs, adjectives, or the like. Our work provides a viable option towards closing this gap, as it makes use of lexical-semantic knowledge covering any part of speech.

Relation disambiguation. The task of disambiguating semantic relations (also called *sense linking* and *relation anchoring*) has been previously described in the context of machine-readable dictionaries (Krovetz, 1992) and ontology learning (Pantel and Pennacchiotti, 2008). Meyer and Gurevych (2010) discussed relation disambiguation for the German Wiktionary using a disambiguation method based on textual similarity. In Section 3.3, we will compare this approach to our system.

The disambiguation of all words in a sense definition (i.e. *gloss disambiguation*), as it has been done in the WordNet 2/eXtendend WordNet project (Harabagiu et al., 1999; Mihalcea and Moldovan, 2001), is very similar to the disambiguation of semantic relations. Therefore, many of the features defined in Section 3.1 are similar to those proposed by Moldovan and Novischi (2004). Note however that we use explicitly defined semantic relations rather than sense definitions as our disambiguation subjects. In addition to that, we adapt our method to Wiktionary instead of using WordNet-specific features and also extend this work to a cross-lingual setting. Very recently, Flati and Navigli (2012) proposed a graph-based method to gloss disambiguation outperforming previous approaches. While this method could in general be adapted to disambiguating relations, we observe that the graph induced by Wiktionary's semantic relations is very sparse. This would hinder finding the cycles and quasi-cycles required by the method.

The disambiguation of translations has been studied in the context of bilingual dictionaries and corpora (Kikui, 1999; Tsunakawa and Kaji, 2010). Mausam et al. (2009) discovered new translations in Wiktionary using a graph-based inference algorithm for Wiktionary translations. Although this also involves a disambiguation of translations, their work is not directly comparable to ours, since they do not strictly use the word senses encoded in Wiktionary but define them based on the translations shared across multiple languages. In contrast to that, we aim at exploiting a wide range of lexical-semantic knowledge and therefore need to rely on the word senses actually encoded in Wiktionary.

[3] `http://www.illc.uva.nl/EuroWordNet/finalresults-ewn.html` (accessed 2012-07-11)

Inference of relations. New semantic relations have been previously inferred when bootstrapping wordnets, i.e. translating the word senses and their definitions to a new language and reusing the relations from an existing wordnet. This has been done, for example, for constructing the Spanish (Atserias and Villarejo, 2004), French (Sagot and Fišer, 2008), and Thai (Thoongsup et al., 2009) wordnets. Such approaches differ from our work in that they do not require a disambiguation of relations. Huang et al. (2002) studied the cross-lingual inference of semantic relations when using imprecise translations. They measure an error rate of 11% for the inference of Chinese semantic relations based on the English WordNet.

3 Disambiguation of Wiktionary's Semantic Relations and Translations

In this section, we describe and evaluate our method for automatically disambiguating semantic relations and translations in Wiktionary.

3.1 Feature Definition

Let $t_j \in t$ be one of multiple possible target word senses for a relation (either a semantic relation or a translation) $r = (s_i, t)$. We define the following features based on our analysis of 200 Wiktionary relations (referred to as *development data*).

Definition overlap. A widely used method for word sense disambiguation is based on counting word overlaps between sense definitions (Lesk, 1986). Let $\text{gloss}(s_i)$ and $\text{gloss}(t_j)$ be the lemmatized and stop-word-filtered sense definitions of s_i and t_j. Their overlap is the number of shared words:

$$f_{\text{Lesk}} := |\text{gloss}(s_i) \cap \text{gloss}(t_j)|.$$

We additionally define f_{ExtLesk} by employing the extension by Banerjee and Pedersen (2003), i.e. we assign squared scores to consecutive sequences of words. If both definitions contain, for example, *"large carnivorous animal"*, we assign a score of $3^2 = 9$.

Source lemma. A special case of overlapping definitions is that the lemma of the source word sense is contained in the definition of the target word sense:

$$f_{\text{src}} := \text{lemma}(s_i) \in \text{gloss}(t_j).$$

This happens frequently, since a definition usually contains synonymous words or follows the *genus-differentia* pattern – i.e., providing a more specialized term (the *genus*) and the properties that distinguish the word from its co-hyponyms (the *differentia*). Consider, for instance, two word senses for *peck*: (i) "[…] *a dry measure of eight quarts"* and (ii) *"a great deal; a large or excessive quantity"*. The second one happens to be the correct disambiguation for the synonymy relation between *deal* and *peck* as it contains the source lemma *deal*.

Linguistic labels. Many word senses are domain-specific, such as the use of *host* as a certain kind of server in computer science. In dictionaries, domain-specific word senses are often marked by linguistic labels stating the domain, register, time, etc. this word sense is normally used in. An example is the sense *"(UK, pejorative) A working-class youth* […]*"* of *chav*. Relations usually connect two word senses of the same domain, register, etc. Hence, we add a feature

$$f_{\text{lbl}} := |\text{label}(s_i) \cap \text{label}(t_j)|$$

counting the number of labels shared by s_i and t_j. Since Wiktionary's linguistic labels are very heterogeneous and fine-grained, we manually grouped similar labels into broader categories; *zoology* and *ornithology* are, for instance, grouped into *biology*.

Inverse relation. Consider a relation between two polysemous words, such as the antonymy relation between $fall_i$ and *increase*. If there is a word sense j of *increase* for which an inverse antonymy relation ($increase_j$, *fall*) is encoded, then it is very likely that j is the correct disambiguation. Let relations(t_j) be the set of related lexical items of t_j. We define

$$f_{\text{inv}} := \text{lemma}(s_i) \in \text{relations}(t_j)$$

as the feature checking for inverse relations.

Relation overlap. The idea of inverse relations can be further extended by finding relations to other words shared by both the source and the target sense. A relation (*sweater*, *cloth*) can, for instance, be disambiguated by finding that one of their word senses shares a relation to *pullover* (a synonym of *sweater* and a hyponym of *cloth*). We define

$$f_{\text{rel}} := \frac{|\text{relations}(s_i) \cap \text{relations}(t_j)|}{|\text{relations}(s_i) \cup \text{relations}(t_j)|},$$

which is similar to the link-based similarity measure proposed by Milne and Witten (2008), who use hyperlinks from Wikipedia.

Commonness and Monosemy. The word senses of a lexicon are often ordered according to their usage frequencies in a corpus or the intuitions of the lexicographers. This has led to a very strong baseline for word sense disambiguation by always choosing the first sense. The same applies to the disambiguation of relations when choosing the first target sense. Therefore, we introduce a feature $f_{\text{idx}} := j$ that is set to the index of the target sense t_j.

Finally, we add a feature f_{mono} that is true if the target word has only one word sense, i.e. if it is monosemous. In these cases, it is most likely that this sense is the correct disambiguation; e.g., for the synonymy relation between *eggplant* and the monosemous word *brinjal*.

Cross-lingual features. Most of the features described above are also applicable in a multilingual setting when using translations instead of semantic relations. In order to also use the features based on sense definitions, we automatically translate them using the Bing translation[4] service. This opens up interesting research opportunities, since the definition of either the source or the target sense can be translated, i.e.

$$f_{\text{Lesk,TL}} := |\text{gloss}(\text{translate}(s_i)) \cap \text{gloss}(t_j)| \text{ or } f_{\text{Lesk,SL}} := |\text{gloss}(s_i) \cap \text{gloss}(\text{translate}(t_j))|.$$

There can even be a combined feature:

$$f_{\text{Lesk,SL\&TL}} := \tfrac{1}{2}(f_{\text{Lesk,SL}} + f_{\text{Lesk,TL}}).$$

Regarding the linguistic label feature f_{lbl}, we manually mapped English and German labels that represent the same meaning (e.g., *biology* and *Biologie*). This yielded a list of 19 label groups covering 1,267 distinct linguistic labels from two languages.

Constraints. In addition to the features introduced above, we can apply a threshold to convert a numeric feature into a boolean one. The notation $f_{\text{Lesk}\geq k}$ defines, for instance, a feature that is true if the sense definitions share at least k words. We use the notation $\hat{f}$ when only the target word sense with the highest feature value is used. The feature $\hat{f}_{\text{Lesk}\geq k}$ is thus true if, and only if, f_{Lesk} is higher than k and the maximum f_{Lesk} of all possible target word senses t.

[4] `http://www.microsofttranslator.com/`

3.2 Disambiguation Method

Let F be a set of features. Based on the notation introduced above, we define a generic relation disambiguating method

$$D\colon\ (r, t_j, F) \mapsto \{0, 1\},$$

returning 1 if t_j is a correct disambiguation for r and 0 otherwise. A basic method $D[f] = f$ uses only a single boolean feature $f \in F$. Thereby, we can model a most frequent sense baseline $MFS = D[f_{\text{idx}=1}]$ always using the first target word sense. One way of combining features is to concatenate them using a backoff strategy, i.e. a method

$$D[f_1 \circ f_2] = \begin{cases} D[f_1] & \text{if } f_1 \in F \\ D[f_2] & \text{otherwise} \end{cases}$$

relying on feature f_1 (if present) and f_2 otherwise. For example, $D[f_{\text{inv}} \circ f_{\text{idx}=1}]$ disambiguates those relations that have an inverse relation using f_{inv}. The remaining relations are disambiguated using a most frequent sense approach.

Based on the features introduced above, we now propose our disambiguation method

$$WKTWSD = D[f_{\text{mono}} \circ f_{\text{lbl}\geq 1} \circ f_{\text{rel}\geq 0.5} \circ f_{\text{src}} \circ f_{\text{inv}} \circ \hat{f}_{\text{ExtLesk}\geq 2} \circ f_{\text{idx}=1}]$$

that concatenates all features introduced above. For the cross-lingual datasets, we use $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL\&TL}}$ instead of $\hat{f}_{\text{ExtLesk}\geq 2}$. The ordering and the thresholds have been chosen based on our analysis of the development data.

3.3 Empirical Evaluation

Comparison to previous work. Our experimental setup is directly comparable to the disambiguation of semantic relations in the German Wiktionary reported by Meyer and Gurevych (2010). They use a publicly available dataset, which consists of 250 manually disambiguated Wiktionary relations. Table 1 shows the performance of our proposed method in comparison with their text-similarity-based method *MG10*. Note that Meyer and Gurevych (2010) evaluated their system by measuring the agreement between the method and each of the two human raters. We therefore report A_O and Cohen's κ (Artstein and Poesio, 2008) following the original experimental setup. The inter-rater agreement serves as an upper bound and the most frequent sense baseline *MFS* is used as a lower bound. Our *WKTWSD* method outperforms their approach by a large margin. The improvement is statistically significant.[5]

Gold standard datasets. To our knowledge, there are no other evaluation datasets for disambiguating Wiktionary relations. That is why we create four new annotated datasets that consist of English semantic relations ($R_{\text{en:en}}$), German semantic relations ($R_{\text{de:de}}$), English–German translations ($R_{\text{en:de}}$), and German–English translations ($R_{\text{de:en}}$). The relations are sampled according to their type, the part of speech, and the number of candidates (i.e., possible target word senses) in order to create a balanced dataset.[6] Balancing out the datasets is very useful for being able to evaluate our approach separately for each sample group and to avoid datasets with a strong bias (e.g., on synonyms between nouns). None of the sampled relations occurs in our development data. Table 2 shows the numbers of sampled relations and the possible target senses (i.e., the number of annotations required).

[5] McNemar's test; $p < .05$

[6] Our sampling procedure is explained in detail in the supplementary material that is published with the datasets.

Method	$A_{O,1}$	$A_{O,2}$	κ_1	κ_2
MFS	.78	.79	.45	.50
MG10	.79	.82	.48	.57
WKTWSD	**.84**	**.85**	**.59**	**.65**
Human	.89	.89	.73	.73

Table 1: Comparison of our system to previous work

	$R_{\text{en:en}}$	$R_{\text{de:de}}$	$R_{\text{en:de}}$	$R_{\text{de:en}}$
Relations	394	459	204	204
Annotations	1,117	1,119	614	656
A_O	.91	.92	.89	.90
κ	.82	.85	.73	.75
F_1	.89	.92	.80	.83

Table 2: Statistics on our evaluation datasets

We then asked two human raters to annotate the monolingual datasets $R_{\text{en:en}}$ and $R_{\text{de:de}}$ and three raters to annotate the cross-lingual datasets $R_{\text{en:de}}$ and $R_{\text{de:en}}$. The raters should annotate each possible target word sense as being a correct ($D = 1$) or incorrect ($D = 0$) disambiguation for the given relation, for example:

s_i = *phenomenal*	D	t_j = *awesome*
(colloquial) Very remarkable; highly extraordinary; amazing.	0	Causing awe or terror; inspiring wonder or excitement.
(colloquial) Very remarkable; highly extraordinary; amazing.	1	(informal) Excellent, exciting, remarkable.

It was allowed to rate all target senses of a relation as incorrect (e.g., if the correct target sense has not yet been encoded in Wiktionary) or to rate more than one target sense as correct (e.g., if the target senses are more fine-grained than the source sense). Each rater was allowed to consult external sources such as lexicons, encyclopedias, etc. (and in particular Wiktionary itself). They were, however, not allowed to contact each other. The raters are native in German and speak English fluently. They have been trained using some example cases and an annotation guidebook that we publish along with the paper.

To estimate the reliability of our datasets, we measure the inter-rater agreement. Table 2 shows the observed agreement A_O and the kappa statistics κ for each dataset. We report Cohen's κ for the two rater case and Fleiss' κ (multi-π) for the three rater case (Artstein and Poesio, 2008). The raters agree on about 90% of the cases. The κ statistics of over .80 for the monolingual datasets suggests good reliability. The cross-lingual datasets have a slightly lower agreement. The disambiguation of translations hence seems to be more difficult for our raters. However, the κ scores are well above .67 and therefore allow us to draw tentative conclusions (Artstein and Poesio, 2008). We also provide F_1 scores for our datasets as suggested by Hripcsak and Rothschild (2005), which serve as upper bounds for our methods.

Finally, we create gold standard datasets based on the majority vote of the raters. As a tie breaker for the monolingual datasets, an additional adjudicator has been asked for a final decision. All datasets including analyses are freely available from our homepage.

Evaluation results. Table 3 shows the performance of our disambiguation method on the four gold standard datasets. We have counted the number of correct decisions $TP + TN$, the number of false positives *FP* and false negatives *FN*, which we use to report accuracy $A = \frac{TP+TN}{N}$, precision $P = \frac{TP}{TP+FP}$ (proportion of correctly disambiguated relations in the system result), recall $R = \frac{TP}{TP+FN}$ (proportion of correctly disambiguated relations in the gold standard), and the $F_1 = \frac{2PR}{P+R}$ score (Manning and Schütze, 1999). As a lower bound, we use the most frequent sense method *MFS*. The upper bound is human performance (*Human*) estimated by the inter-rater agreement A_O and the inter-rater F_1 score introduced above. Our *WKTWSD* method significantly outperforms the *MFS* baseline for each dataset. The only exception being the precision on the $R_{\text{de:en}}$ dataset, which is slightly lower than the precision of *MFS*.

Besides the lower and upper boundaries, we trained a number of machine learning classifiers for our set of features using the Weka toolkit (Hall et al., 2009). We report the results for a Naïve Bayes (*NB*) and a *J48* decision tree (a C4.5 clone) here, although we tried other classifiers as well, which generally yielded similar results. The training was done in a 5-fold cross validation. Note that we did not optimize the configuration in order to avoid overfitting to the datasets. In general, our *WKTWSD* method reaches a similar or even better performance than the machine learning classifiers. The main reason for this is the largely varying number of possible target word senses. While one relation might have only a single possible target sense, another one might have ten or even more. This tends to cause more false negatives in the machine learning methods and thus less relations that can be disambiguated. The finding is in line with previous work on gloss disambiguation: Moldovan and Novischi (2004) note that compiling a sufficient set of training examples is not possible in many cases. Despite this, the machine learning methods mostly achieve a slightly higher precision. *J48* even yields $P = .82$ for the $R_{\text{de:en}}$ dataset. However, this always comes at the cost of a lower recall.

Feature and error analysis. Table 4 shows the precision P and coverage C (proportion of items covered by this feature) of using each feature $f \in F$ individually. With the exception of $f_{\text{idx=1}}$ (most frequent sense strategy), none of the features is able to disambiguate the whole dataset, but most of them achieve a very high precision on the covered items. It is not surprising that f_{mono} performs extremely well ($P \in [.88, .96]$), since there is only one target word sense available for these cases. The feature f_{src} performs well on the monolingual datasets ($P \in [.87, .97]$), but does not work at all on the cross-lingual task ($P \in [.38, .50]$). The reasons for this are ambiguities in the sense definitions that are often not resolved by the machine translation service. Parallel ambiguities such as *commission* and *Kommission*, which both mean either a group of people or a transaction fee of a broker, is a main source of errors here. Similar errors also occur for f_{inv}. The word overlap feature $\hat{f}_{\text{ExtLesk}}$ generally shows a high precision. It is, in particular, higher than usually reported for word sense disambiguation tasks (Navigli, 2009). The reason might be that we do not compare a sense definition with context words, but two definitions with each other and hence benefit from comparing texts that are specially crafted to characterize word senses. Interestingly, the imprecise translation of certain words noted for f_{src} is less problematic for $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL\&TL}}$, as there are usually at least some correctly translated words in the sense definition. In our experiments, we found that $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL}}$ outperforms $\hat{f}_{\text{ExtLesk}\geq 2,\text{TL}}$, whereas $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL\&TL}}$ is only marginally better than $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL}}$. The English Wiktionary is very sparse in encoding semantic relations. The coverage of $f_{\text{rel}\geq 0.5}$ is therefore very low for all datasets involving English data.

Since we ordered the features manually for our *WKTWSD* method, we additionally define a method *BestOrder* which concatenates the features in descending order of their precision on

Method	$R_{\text{en:en}}$ A	P	R	F_1	$R_{\text{de:de}}$ A	P	R	F_1
MFS	.81	.75	.74	.74	.79	.78	.76	.77
WKTWSD	.84	.78	**.80**	**.79**	.84	.83	**.83**	**.83**
NB	.85	**.81**	.78	**.79**	.84	**.84**	.81	.82
J48	.83	**.81**	.71	.76	.84	.83	.82	**.83**
BestOrder	.85	.79	.80	.80	.85	.84	.83	.84
Human	.91			.89	.92			.92

Method	$R_{\text{en:de}}$ A	P	R	F_1	$R_{\text{de:en}}$ A	P	R	F_1
MFS	.79	.62	.72	.67	.79	.64	.66	.65
WKTWSD	.81	.64	**.75**	**.69**	.79	.62	**.71**	**.67**
NB	.81	.67	.69	.68	.82	.74	.61	**.67**
J48	.79	**.69**	.53	.60	.82	**.82**	.53	.64
BestOrder	.80	.63	.75	.69	.81	.67	.73	.70
Human	.89			.80	.90			.83

Table 3: Performance of our disambiguation methods on the four evaluation datasets

Feature	$R_{\text{en:en}}$ P	C	$R_{\text{de:de}}$ P	C	$R_{\text{en:de}}$ P	C	$R_{\text{de:en}}$ P	C
f_{mono}	.91	.21	.94	.22	.96	.08	.88	.08
f_{inv}	.78	.13	.89	.31	.68	.49	.67	.41
$f_{\text{lbl}\geq 1}$	.82	.07	.90	.05	.86	.02	.60	.04
f_{src}	.87	.10	.97	.07	.50	.20	.38	.18
$f_{\text{rel}\geq 0.5}$	.94	.04	.90	.14	.33	.01	.75	.01
$\hat{f}_{\text{ExtLesk}\geq 2}$	.89	.27	.99	.12	.87	.15	.93	.17
$f_{\text{idx}=1}$	.75	1.0	.78	1.0	.62	1.0	.64	1.0

Table 4: Precision and coverage of each feature

each dataset. The rationale behind this is that we make use of the best feature before moving to the next one. By comparing *WKTWSD* to *BestOrder*, we can measure the influence of our manually chosen ordering. Note, however, that *BestOrder* needs to be considered as an upper bound for *WKTWSD* rather than a separate method, because it made use of our analysis of the test data. The results can be found in Table 3. We observe that the order of the features plays only a minor role: *WKTWSD* and *BestOrder* are only slightly different although they concatenate the features in totally different ways. The largest difference accounts to .03 for the $R_{\text{de:en}}$ dataset and is mostly due to the low performance of f_{src}.

Summary. We conclude that our approach is better suited for disambiguating Wiktionary relations than previous works using textual similarity. The features are effectively applied using a concatenation method. The training of machine learning classifiers could not improve these results in our experiments.

	Our resource		Wordnets	
	English	German	WordNet	GermaNet
Lexical entries	379,694	85,574	156,584	85,257
Word senses	474,128	73,500	206,978	96,690
Semantic relations	215,353	300,724	1,398,868	512,653
…Synonyms	*70,199*	*78,133*	*315,984*	*74,552*
…Antonyms	*35,291*	*33,391*	*7,979*	*3,359*
…Hypo-/Hypernyms	*54,494*	*87,246*	*658,804*	*397,335*
…Other types	*55,269*	*101,954*	*416,101*	*37,407*
Translations	79,382		16,347	

Table 5: Statistics on our new resource in comparison to WordNet and GermaNet

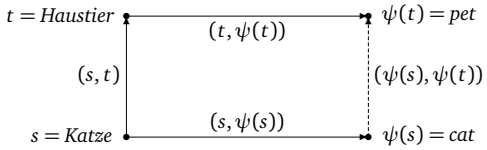

Figure 2: Cross-lingual inference of the semantic relation $(\psi(s), \psi(t))$

4 A Multilingual, Sense-Disambiguated Wiktionary

Resource construction. We create our new multilingual resource by using all word senses encoded in a given set of Wiktionary language editions (English and German in our experiments). Then, we perform the automatic disambiguation of the semantic relations and translations to obtain a fully disambiguated resource. We use the *WKTWSD* method for this task, as it performed well on our evaluation datasets. The disambiguated translations allow for extracting lexical-semantic knowledge in multiple languages. The first sense of *(to) stroll* is, for instance, *"to wander on foot* […]*"* in the English Wiktionary. When following its translations, we are able to extract the German equivalent *spazieren*: *"gemächlich gehen* […]*"*. In this way, we can also obtain multilingual example sentences, linguistic labels, etc.

Inference of relations. For semantic relations, we can even further benefit from their disambiguated target senses: Let (s, t) be a disambiguated semantic relation in one of the Wiktionary language editions and let $(s, \psi(s))$ and $(t, \psi(t))$ be disambiguated translations of s and t into another language. Assuming a correct disambiguation of these three relations, we can infer a fourth relation $(\psi(s), \psi(t))$, since the meaning of s and t is preserved under the disambiguated translations. Figure 2 shows an example: For the German hypernym (*Katze*, *Haustier*) and the corresponding translations (*Katze, cat*) and (*Haustier, pet*), we can infer the English hypernymy relation (*cat, pet*) that is currently not encoded in the English Wiktionary. Note that the inferred relation is also sense-disambiguated, i.e. both *cat* and *pet* refer to the animal sense.

Size of our resource. Our final resource contains 215,353 English and 300,724 German semantic relations. The English Wiktionary benefits most from inferring new semantic relations: We increased the number of relations found in the original Wiktionary (26,965) by almost an order of magnitude. But also for the German language, we were able to infer 10,705 new semantic relations. In addition to that, our resource consists of 474,128 English and

73,500 German word senses as well as 79,382 translations (45,246 English–German and 34,136 German–English). Table 5 shows detailed statistics of our resource including the most common types of the encoded semantic relations. We compare our resource with the Princeton WordNet (Fellbaum, 1998) and GermaNet (Kunze and Lemnitzer, 2002) and their inter-lingual index (which is a part of EuroWordNet). Our resource surpasses the number of translations by a large margin, but contains less semantic relations than in the expert-built wordnets. The coverage of lexical entries and word senses is comparable or higher.

5 Measuring Verb Similarity

To demonstrate the usefulness of our resource, we carry out two experiments employing the newly created resource in a monolingual and cross-lingual verb similarity task. Judging verb similarity is of particular interest for applications such as cross-lingual word sense disambiguation (Lefever and Hoste, 2010), lexical substitution (Mihalcea et al., 2010), or question answering (Magnini et al., 2005). State of the art knowledge-based systems rely heavily on Wikipedia, which predominantly encodes encyclopedic knowledge about nouns. The large amount of multilingual lexical-semantic knowledge in Wiktionary let us expect good results not only for nouns, but also for other parts of speech and verbs in particular.

Monolingual verb similarity. Yang and Powers (2006) introduced an evaluation dataset for verb similarity that consists of 130 English verb pairs taken from TOEFL and ESL (English as a second language) questions. For each of them, a numerical score is provided expressing the human intuitions of their similarity. These scores are averaged over six human annotators that were asked to rate the similarity of each pair on a graded scale from 0 (*not at all related*) to 4 (*inseparably related*). Yang and Powers (2006) report a correlation of $r = 0.866$ between the raters. An example from their dataset is the verb pair (*approve, support*) with a score of 3.

To the best of our knowledge, Zesch et al. (2008b) reports the latest evaluation results on this dataset as shown in column *Z08* of Table 6. They use explicit semantic analysis, a method based on concept vectors (Gabrilovich and Markovitch, 2007) built from WordNet, Wikipedia, and the undisambiguated Wiktionary. Each entry from these resources (synsets in WordNet and wiki pages in Wikipedia and Wiktionary) is regarded as one concept. For a given word pair, two concept vectors are then created that consist of the word's tf-idf scores over the concepts. The similarity for this word pair is then expressed by the cosine of the two concept vectors. Although Zesch et al. (2008b) find Wiktionary to yield best results for computing semantic relatedness between nouns, the performance for verb similarity is substantially lower than using WordNet. One reason for that is the high degree of polysemy of verbs, which is not dealt with by their approach. Since our resource is completely sense-disambiguated, we can, in contrast, compute sense-disambiguated concept vectors using each word sense as one concept.

We reproduced the results of Zesch et al. (2008b), also using WordNet, Wikipedia, and the undisambiguated Wiktionary, and show them in the column $V_{\text{en:en}}$ of Table 6. Note that we use all 130 verb pairs, whereas Zesch et al. (2008b) used only the 80 pairs that were covered by all three similarity metrics they tried. Therefore, our scores slightly differ from *Z08*. In addition to the three resources, we report the performance when using the sense-disambiguated concept vectors derived from our resource. Using our resource yields better results than using Wikipedia or the undisambiguated Wiktionary. The previously best resource WordNet is slightly outperformed by our resource. This difference is, however, not statistically significant. All four concept-vector-based methods cover 100% of the dataset and are thus directly comparable.

Resource	*Z08*	$V_{\text{en:en}}$	$V_{\text{de:de}}$	$V_{\text{en:de}}$	$V_{\text{de:en}}$
WN/GN	**.71**	.69	**.57**	.31	.23
Wikipedia	.29	.27	.33	.23	.28
Wiktionary	.65	.63	.36	—	—
Our resource	—	**.73**	.52	**.53**	**.51**
Coverage	62%	100%	92%	95%	97%

Table 6: Evaluation results on the four verb similarity datasets using concept vectors from WordNet/GermaNet (WN/GN), Wikipedia, Wiktionary, and our new resource in comparison to previous work by Zesch et al. (2008b) (*Z08*). Performance is measured by Spearman's rank correlation coefficient using Horn's correction for tied ranks (Horn, 1942). All correlations significantly differ from random (two-tailed paired t-test; $p < .05$).

We also study German verb similarity and therefore translate the $V_{\text{en:en}}$ dataset. The verb pair (*approve, support*) is, for instance, translated to (*annehmen, unterstützen*) keeping its similarity score of 3. Table 6 shows the results for this new $V_{\text{de:de}}$ dataset. To create the concept vectors, we use GermaNet instead of WordNet as well as the German editions of Wikipedia and Wiktionary. We use only the 120 verb pairs covered by all four resources. Our resource is again able to outperform Wikipedia and the undisambiguated Wiktionary by a wide margin. The performance competes with the expert-built GermaNet, but is slightly lower than that. As opposed to the English language, GermaNet and the German part of our resource are similar in size (see Table 5), which can explain these results. This is why we expect better results with the growth of the German Wiktionary. Furthermore, we conclude that our resource can be a promising alternative for languages with less developed expert-built resources.

Cross-lingual verb similarity. Based on the English and the German verb pairs, we create two cross-lingual verb similarity datasets that use the first English verb together with the second German verb from each corresponding verb pair $V_{\text{en:de}}$ and, vice versa, the first German verb together with the second English verb $V_{\text{de:en}}$. For the example introduced above, this yields the two verb pairs (*approve, unterstützen*) and (*annehmen, support*), both with a score of 3.

Table 6 shows the evaluation results using these two datasets. To create the cross-lingual concept vectors, we use the inter-lingual index between WordNet and GermaNet, the interwiki links from Wikipedia, and the disambiguated translations from our new resource. Since the translations of the original Wiktionary are not sense-disambiguated, they cannot be used to build cross-lingual concept vectors.[7] As noted in Section 2, the inter-lingual index of WordNet and GermaNet (which is part of EuroWordNet) is very small. Consequently, we observe that the expert-built wordnets yield a substantially lower performance for $V_{\text{en:de}}$ and $V_{\text{de:en}}$ than in the monolingual setting. Wikipedia likewise yields low scores because of its lack of the knowledge about verbs, whereas our resource significantly outperforms ($p < .01$) both the expert-built wordnets and Wikipedia.

Our error analysis shows that many of the judgments derived from our resource are useful. The predominant problem is still the coverage of the translations. The similarity of the English–German verb pair *(concoct, ausarbeiten)* is, for instance, not yet backed up by a translation in Wiktionary and is hence underestimated by the system. While this is essentially the same problem as for the wordnets, the problem is much less severe for our resource.

[7] Wiktionary also encodes interwiki links for each wiki page, but they link to the same form (e.g. from *walk* in the English Wiktionary to *walk* in the German Wiktionary) rather than to translations and thus cannot be used.

Summary. Wikipedia-based resources are not very appropriate for computing verb similarity as they focus on encyclopedic knowledge about nouns. Expert-built wordnets work well for computing monolingual verb similarity, because they have a sufficient coverage and encode thoroughly elaborated lexical-semantic knowledge. Our new disambiguated Wiktionary-based resource competes with their quality. Since Wiktionary is available in over 170 languages, our approach is, however, also applicable to the languages lacking large expert-built resources. In a cross-lingual setting, this shows a different picture: Expert-built multilingual wordnets suffer from their small size. Since the disambiguated translations in our resource let us build cross-lingual concept vectors, they can be effectively utilized in this task.

Conclusion and perspectives

We have created a new multilingual, sense-disambiguated resource using the word senses from Wiktionary and interconnecting them by means of disambiguated semantic relations and translations. For the automatic disambiguation of the relations, we proposed and evaluated a rule-based method using seven different features. Our features are similar to those used by Moldovan and Novischi (2004), whereas we adjusted them to our specific task and generalized them to the cross-lingual setting. We found our method to significantly outperform a previous approach based on textual similarity (Meyer and Gurevych, 2010). In a second evaluation based on four newly created datasets, we obtained promising results exceeding the baseline in every case. Using the disambiguated relations, we inferred a large number of new semantic relations and thereby yielded almost a tenfold increase in the number of relations for the English language. Our final resource fills the gap between small expert-built multilingual wordnets and Wikipedia-based resources, which are mostly restricted to the encyclopedic knowledge about nouns. The new resource and all evaluation data is publicly available for research.[8]

We also employed our new resource in a monolingual and cross-lingual verb similarity task. Besides the standard dataset by Yang and Powers (2006), we created a novel German and two cross-lingual verb similarity datasets. Our resource competes with expert-built wordnets in the monolingual setting. Since Wiktionary is available in many languages, this allows for computing verb similarity also for languages lacking large expert-built resources. In the cross-lingual setting, our sense-disambiguated resource outperforms both Wikipedia and the expert-built wordnets. The former suffers from the small amount of knowledge about verbs and the latter lack coverage of the inter-lingual index.

In future work, we plan to combine our resource with BabelNet or UBY (Gurevych et al., 2012) in order to benefit from the heterogeneous knowledge found in WordNet, Wikipedia, and our resource. Extending our resource to other languages and exploring alternative disambiguation algorithms such as CQC are further promising options. We will also consider providing our inferred semantic relations to the Wiktionary community to contribute to the harmonization of Wiktionary data. Besides verb similarity, our sense-disambiguated resource has the potential to improve other natural language processing tasks as well, for instance, question answering.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Christian Kirschner, Dr. Judith Eckle-Kohler, and Dr. Torsten Zesch for their contributions to this project as well as Dongqiang Yang and David M. W. Powers for sharing their verb similarity dataset.

[8] http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/

References

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Atserias, J. and Villarejo, Luís Rigau, G. (2004). Spanish WordNet 1.6: Porting the Spanish WordNet Across Princeton Versions. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 161–164, Lisbon, Portugal.

Banerjee, S. and Pedersen, T. (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, Mexico.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia – A Crystallization Point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

de Melo, G. and Weikum, G. (2009). Towards a Universal Wordnet by Learning from Combined Evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 513–522, Hong Kong, China.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. Cambridge, MA: MIT Press.

Flati, T. and Navigli, R. (2012). The CQC Algorithm: Cycling in Graphs to Semantically Enrich and Enhance a Bilingual Dictionary. *Journal of Artificial Intelligence Research*, 43:135–171.

Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, Hyderabad, India.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY – A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Harabagiu, S. M., Miller, G. A., and Moldovan, D. I. (1999). WordNet 2 - A Morphologically and Semantically Enhanced Resource. In *Proceedings of the ACL Special Interest Group on the Lexicon Workshop on Standardizing Lexical Resources*, pages 1–7, College Park, MD, USA.

Herbert, B., Szarvas, G., and Gurevych, I. (2011). Combining Query Translation Techniques to Improve Cross-Language Information Retrieval. In Clough, P., Foley, C., Gurrin, C., Jones, G. J., Kraaij, W., Lee, H., and Murdoch, V., editors, *Advances in Information Retrieval: 33rd European Conference on IR Research*, volume 6611 of *Lecture Notes in Computer Science*, pages 712–715. Berlin/Heidelberg: Springer.

Horn, D. (1942). A Correction for the Effect of Tied Ranks on the Value of the Rank Difference Correlation Coefficient. *Journal of Educational Psychology*, 33(9):686–690.

Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Huang, C.-R., Tseng, I.-J. E., and Tsai, D. B. S. (2002). Translating lexical semantic relations: the first step towards multilingual wordnets. In *Proceedings of the COLING '02 Workshop on 'Building and Using Semantic Networks'*, Taipei, Taiwan.

Kikui, G. (1999). Resolving Translation Ambiguity Using Non-parallel Bilingual Corpora. In *Proceedings of the ACL '99 Workshop on 'Unsupervised Learning in Natural Language Processing'*, pages 31–36, College Park, MD, USA.

Krovetz, R. (1992). Sense-Linking in a Machine Readable Dictionary. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 330–332, Newark, DE, USA.

Kunze, C. and Lemnitzer, L. (2002). GermaNet — representation, visualization, application. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, volume 5, pages 1485–1491, Las Palmas, Canary Islands, Spain.

Lefever, E. and Hoste, V. (2010). SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden.

Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, Toronto, ON, Canada.

Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., and Sutcliffe, R. (2005). Overview of the CLEF 2004 Multilingual Question Answering Track. In Peters, C., Clough, P., Gonzalo, J., Jones, G. J., Kluck, M., and Magnini, B., editors, *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum,*, volume 3491 of *Lecture Notes in Computer Science*, pages 371–391. Berlin/Heidelberg: Springer.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Mausam, Soderland, S., Etzioni, O., Weld, D., Skinner, M., and Bilmes, J. (2009). Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 262–270, Singapore.

Medelyan, O., Legg, C., Milne, D., and Witten, I. H. (2009). Mining Meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

Meyer, C. M. and Gurevych, I. (2010). Worth its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing: 11th International Conference*, volume 6008 of *Lecture Notes in Computer Science*, pages 38–49. Berlin/Heidelberg: Springer.

Mihalcea, R. and Moldovan, D. (2001). eXtended WordNet: Progress Report. In *Proceedings of the NAACL '01 Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations'*, pages 95–100, Pittsburgh, PA, USA.

Mihalcea, R., Sinha, R., and McCarthy, D. (2010). SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations*, pages 9–14, Uppsala, Sweden.

Milne, D. and Witten, I. H. (2008). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *Proceedings of the AAAI '08 Workshop 'Wikipedia and Artificial Intelligence: An Evolving Synergy'*, pages 25–30, Chicago, IL, USA.

Moldovan, D. and Novischi, A. (2004). Word sense disambiguation of WordNet glosses. *Computer Speech and Language*, 18(3):301–317.

Nastase, V., Strube, M., Börschinger, B., Zirn, C., and Elghafari, A. (2010). WikiNet: A very large scale multi-lingual concept network. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1015–1022, Valetta, Malta.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Neff, M. S. and McCord, M. C. (1990). Acquiring Lexical Data from Machine-Readable Dictionary Resources for Machine Translation. In *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, pages 85–90, Austin, TX, USA.

Oepen, S., Velldal, E., Lønning, J. T., Meurer, P., Rosén, V., and Flickinger, D. (2007). Towards Hybrid Quality-Oriented Machine Translation: On Linguistics and Probabilities in MT. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 144–153, Skövde, Sweden.

Pantel, P. and Pennacchiotti, M. (2008). Automatically Harvesting and Ontologizing Semantic Relations. In Buitelaar, P. and Cimiano, P., editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 171–198. Amsterdam: IOS Press.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the First International WordNet Conference*, pages 293–302, Mysore, India.

Sagot, B. and Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *Proceedings of the LREC '08 Workshop 'Ontologies and Lexical Resources'*, pages 14–19, Marrakech, Morocco.

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiş, D., Koeva, S., Totkov, G., Dutoit, D., and Grigoriadou, M. (2002). BALKANET: A Multilingual Semantic Network for the Balkan Languages. In *Proceedings of the First International WordNet Conference*, pages 12–14, Mysore, India.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International World Wide Web Conference*, pages 697–706, Banff, AB, Canada.

Thoongsup, S., Charoenporn, T., Robkop, K., Sinthurahat, T., Mokarat, C., Sornlertlamvanich, V., and Isahara, H. (2009). Thai WordNet Construction. In *Proceedings of the 7th ACL/IJCNLP '09 Workshop on 'Asian Language Resources'*, pages 139–144, Singapore.

Tsunakawa, T. and Kaji, H. (2010). Augmenting a Bilingual Lexicon with Information for Word Translation Disambiguation. In *Proceedings of the Eighth COLING '10 Workshop on 'Asian Language Resources'*, pages 30–37, Beijing, China.

Tufiş, D., Ion, R., Barbu, E., and Barbu, V. (2004). Cross-Lingual Validation of Multilingual Wordnets. In *Proceedings of the Second Global WordNet Conference*, pages 332–340, Brno, Czech Republic.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*, 32(2–3):73–89.

Yang, D. and Powers, D. M. W. (2006). Verb Similarity on the Taxonomy of WordNet. In *Proceedings of the Third International WordNet Conference*, pages 121–128, Jeju Island, Korea.

Zesch, T., Müller, C., and Gurevych, I. (2008a). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1646–1652, Marrakech, Morocco.

Zesch, T., Müller, C., and Gurevych, I. (2008b). Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 861–867, Chicago, IL, USA.