# On the Utility of Automatically Generated Wordnets

Gerard de Melo and Gerhard Weikum

Max Planck Institute for Informatics
Campus E1 4
66123 Saarbrücken, Germany
{demelo,weikum}@mpi-inf.mpg.de

**Abstract.** Lexical resources modelled after the original Princeton WordNet are being compiled for a considerable number of languages, however most have yet to reach a comparable level of coverage. In this paper, we show that automatically built wordnets, created from an existing wordnet in conjunction with translation dictionaries, are a suitable alternative for many applications, despite the errors introduced by the automatic building procedure. Apart from analysing the resources directly, we conducted tests on semantic relatedness assessment and cross-lingual text classification with very promising results.

## 1 Introduction

One of the main requirements for domain-independent lexical knowledge bases, apart from an appropriate data model, is a satisfactory level of coverage. WordNet is the most well-known and most widely used lexical database for English natural language processing, and is the fruit of over 20 years of manual work carried out at Princeton University [1]. The original WordNet has inspired the creation of a considerable number of similarly-structured resources for other languages ("wordnets"), however, compared to the original, many of these still exhibit a rather low level of coverage due to the laborious compilation process. In this paper, we argue that, depending on the particular task being pursued, one can instead often rely on machine-generated wordnets, created with translation dictionaries from an existing wordnet such as the original WordNet.

The remainder of this paper is laid out as follows. In Section 2 we provide an overview of strategies for building wordnets automatically, focusing in particular on a recent machine learning approach. Section 3 then evaluates the quality of a German wordnet built using this technique, examining the accuracy, coverage, as well as the general appropriateness of automatic approaches. This is followed by further investigations motivated by more pragmatic considerations. After considering human consultation in Section 4, we proceed to look more closely at possible computational applications, discussing our results in monolingual tasks such as semantic relatedness estimation in Section 5, and multilingual ones such as cross-lingual text classification in Section 6. We conclude with final remarks and an exploration of future research directions in Section 7.

## 2 Building Wordnets

In this section, we summarize some of the possible techniques for automatically creating wordnets fully aligned to an existing wordnet. We do not consider the so-called merge model, which normally requires some pre-existing wordnet-like thesaurus for the new language, and instead focus on the expand model, which mainly relies on translations [2]. The general approach is as follows: (1) Take an existing wordnet for some language $L_0$, usually Princeton WordNet for English (2) For each sense $s$ listed by the wordnet, translate all the terms associated with $s$ from $L_0$ to a new language $L_{\mathrm{N}}$ using a translation dictionary (3) Additionally retain all appropriate semantic relations between senses in order to arrive at a new wordnet for $L_{\mathrm{N}}$.

The main challenge lies in determining which translations are appropriate for which senses. A dictionary translating an $L_0$-term $e$ to an $L_{\mathrm{N}}$-term $t$ does not imply that $t$ applies to all senses of $e$. For example, considering the translation of English "`bank`" to German "`Bank`", we can observe that the English term can also be used for riverbanks, while the German "`Bank`" cannot (and likewise, German "`Bank`" can also refer to a park bench, which does not hold for the English term).

In order to address these problems, several different heuristics have been proposed. Okumura and Hovy [3] linked a Japanese lexicon to an ontology based on WordNet synsets. They considered four different strategies: (1) simple heuristics based on how polysemous the terms are with respect to the number of translations and with respect to the number of WordNet synsets (2) checking whether one ontology concept is linked to *all* of the English translations of the Japanese term (3) compatibility of verb argument structure (4) degree of overlap between terms in English example sentences and translated Japanese example sentences.

Another important line of research starting with Rigau and Agirre [4], and extended by Atserias et al. [5] resulted in automatic techniques for creating preliminary versions of the Spanish WordNet and later also the Catalan WordNet [6]. Several heuristic decision criteria were used in order to identify suitable translations, e.g. monosemy/polysemy heuristics, checking for senses with multiple terms having the same $L_{\mathrm{N}}$-translation, as well as heuristics based on conceptual distance measures. Later, these were combined with additional Hungarian-specific heuristics to create a Hungarian nominal WordNet [7].

Pianta et al. [8] used similar ideas to produce a ranking of candidate synsets. In their work, the ranking was not used to automatically generate a wordnet but merely as an aid to human lexicographers that allowed them to work at faster pace. This approach was later also adopted for the Hebrew WordNet [9].

A more advanced approach that requires only minimal human work lies in using machine learning algorithms to identify more subtle decision rules that can rely on a number of different heuristic scores with different thresholds. We will briefly summarize our approach [10]. A classifier $f$ is trained on labelled examples $(x_i, y_i)$ for pairs $(t_i, s_i)$, where $t_i$ is an $L_{\mathrm{N}}$-term and $s_i$ is a candidate sense for $t_i$. Each labelled instance consists of a real-valued feature vector $x_i$, and an indicator $y_i \in \mathcal{Y} = \{0, 1\}$, where 1 denotes a positive example, which implies that linking $t_i$ with sense $s_i$ is appropriate, and 0 characterizes negative examples. Based on these training examples, $f$ classifies new unseen test instances by computing a confidence value $y \in [0,1]$ that indicates to what degree an association is predicted to be correct. One may then obtain a confidence value $y_{t,s}$ for each possible pair $(t,s)$ where $t$ is a $L_{\mathrm{N}}$-term translated to an $L_0$-term that is in turn linked to a sense $s$. These values can be used to create the new wordnet by either maintaining all $y_{t,s}$ as weights in order to create a weighted wordnet, or alternatively one can use confidence thresholds to obtain a regular unweighted wordnet. For the latter case, we use two thresholds $\alpha_1$, $\alpha_2$, and accept a pair $(t,s)$ if $y_{t,s} \geq \alpha_1$, or alternatively if $\alpha_1 > y_{t,s} \geq \alpha_2$ and $y_{t,s} > y_{t,s'}$ for all $s' \neq s$.

The feature vectors $x_i$ are created by computing a variety of scores based on statistical properties of the $(t,s)$ pair as feature values. We mainly rely on a multitude of semantic overlap scores reflecting the idea that senses with a high semantic proximity to other candidate senses are more likely to be appropriate, as well as polysemy scores that reflect the idea that a sense becomes more important when there are few relevant alternative senses. The former are computed as

$$\sum_{e \in \phi(t)} \max_{s' \in \sigma(e)} \gamma(t,s')\ \mathrm{rel}(s,s') \tag{1}$$

while for the latter we use

$$\sum_{e \in \phi(t)} \frac{\mathbf{1}_{\sigma(e)}(s)}{1 + \sum\limits_{s' \in \sigma(e)} \gamma(t,s')(1 - \mathrm{rel}(s,s'))}. \tag{2}$$

In these formulae, $\phi(t)$ yields the set of translations of $t$, $\sigma(e)$ yields the set of senses of $e$, $\gamma(t,s)$ is a weighting function, and $\mathrm{rel}(s,s')$ is a semantic relatedness function between senses. The characteristic function $\mathbf{1}_{\sigma(e)}(s)$ yields 1 if $s \in \sigma(e)$ and 0 otherwise. We use a number of different weighting functions $\gamma(t,s)$ that take into account lexical category compatibility, corpus frequency information, etc., as well as multiple relatedness functions $\mathrm{rel}(s,s')$ based on gloss similarity and graph distance (cf. Section 5.2).

This approach has several advantages over previous proposals: (1) Apart from the translation dictionary, it does not rely on additional resources such as monolingual dictionaries with field descriptors, verb argument structure information, and the like for the target language $L_{\mathrm{N}}$, and thus can be used in many settings, (2) the learning algorithm can exploit real-valued heuristic scores rather than just predetermined binary decision criteria, leading to a greater coverage, (3) the algorithm can take into account complex dependencies between multiple scores rather than just single heuristics or combinations of two heuristics.

## 3 Analysis of a Machine-Generated Wordnet

In the remainder of this paper, we will focus on a German-language wordnet produced using the machine learning technique described above, as it is the most advanced approach. The wordnet was generated from Princeton WordNet 3.0 and the Ding translation dictionary [11] using a linear kernel support vector machine [12] with posterior probability estimation as implemented in LIBSVM [13]. The training set consisted of 1834 candidate mappings for 350 randomly selected German terms that were manually classified as correct (22%) or incorrect. The values $\alpha_1 = 0.5$ and $\alpha_2 = 0.45$ were chosen as classification thresholds.

### 3.1 Accuracy and Coverage

In order to evaluate the quality of the wordnet generated in this manner, we considered a test set of term-sense mappings for 350 further randomly selected terms. We then determined whether the resulting 1624 mappings, which had not been involved in the wordnet building process, corresponded with the entries of our new wordnet. Table 1 summarizes the results, showing the precision and recall with respect to this test set.

**Table 1.** Evaluation of precision and recall on an independent test set

| | precision | recall |
|---|---|---|
| nouns | 79.87 | 69.40 |
| verbs | 91.43 | 57.14 |
| adjectives | 78.46 | 62.96 |
| adverbs | 81.81 | 60.00 |
| overall | 81.11 | 65.37 |

The results demonstrate that indeed a surprisingly high level of precision and recall can be obtained with fully automated techniques, considering the difficulty of the task. While the precision might not fulfil the high lexicographical standards adopted by traditional dictionary publishers, we shall later see that it suffices for many practical applications. Furthermore, one of course may obtain a higher level of precision at the expense of a lower recall by adjusting the acceptance thresholds. For very high recall levels, an increased precision might not be realistic even using purely manual work, considering that Miháltz and Prószéky [7] report an inter-annotator agreement of 84.73% for such mappings.

**Table 2.** Quantitative Assessment of Coverage of the German wordnet

| | sense mappings | terms | lexicalized synsets |
|---|---|---|---|
| nouns | 53146 | 35089 | 28007 |
| verbs | 13875 | 5908 | 6304 |
| adjectives | 21799 | 13772 | 9949 |
| adverbs | 4243 | 2992 | 2593 |
| total | 93063 | 55522 | 46853 |

Table 2 shows that applying the classification thresholds to all terms in the dictionary leads to a wordnet with a considerable coverage. While smaller than GermaNet 5.0 [14], one of the largest wordnets, it covers more senses than any of the original eight wordnets delivered by the EuroWordNet project [2]. Table 3 gives an overview of the polysemy of the terms as covered by our wordnet, with arithmetic means computed from the polysemy either of all terms, or exclusively from terms polysemous with respect to the wordnet.

**Table 3.** Polysemy of Terms and Mean Number of Lexicalizations (excluding unlexicalized senses)

| | mean term polysemy | mean term polysemy excluding monosemous | mean no. of sense lexicalizations |
|---|---|---|---|
| nouns | 1.51 | 2.95 | 1.90 |
| verbs | 2.35 | 4.36 | 2.20 |
| adjectives | 1.58 | 2.79 | 2.19 |
| adverbs | 1.42 | 2.52 | 1.64 |
| total | 1.68 | 3.07 | 1.99 |

A more qualitative assessment of the accuracy and coverage revealed the following issues:

- Non-Uniformity of Coverage: While even many specialized terms are included (e.g. "`Kokarde`", "`Vasokonstriktion`"), certain very common terms were found to be missing (e.g. "`Kofferraum`", "`Schloss`"). This seems to arise from the fact that common terms tend to be more polysemous, though frequently such terms also have multiple translations, which tends to facilitate the mapping process. One solution would be manually adding mappings for terms with high corpus frequency values, which due to Zipf's law would quickly improve the relative coverage of the terms in ordinary texts.
- Lexical Gaps and Incongruencies: Another issue is the lack of terms for which there are no lexicalized translations in the English language, or which are not covered by the source wordnet, e.g. the German word "`Feierabend`" means the finishing time of the daily working hours. The solution could consist in smartly adding new senses to the sense hierarchy based on paraphrasing translations (e.g. as a hyponym of "`time`" for our current example).
- Multi-word expressions in $L_{\mathrm{N}}$: Certain multi-word translations in $L_{\mathrm{N}}$ might be considered inappropriate for inclusion in a lexical resource, e.g. the Ding dictionary lists "`Jahr zwischen Schule und Universität`" as a translation of "`gap year`". By generally excluding all multi-word expressions one would also likely drop a lot of lexicalized expressions, e.g. German "`runde Klammer`" (parenthesis). A much better solution is to automatically mark all multi-word expressions as possibly unlexicalized whenever no matching entry is found in monolingual dictionaries.

### 3.2 Relational Coverage

By producing mappings to senses of an existing source wordnet, we have the great advantage of immediately being able to import relations between those synsets. An excerpt of some of the relations we imported is given in Table 4.

**Table 4.** An excerpt of some of the imported relations. We distinguish full links between two senses both with $L_\mathrm{N}$-lexicalizations, and outgoing links from senses with an $L_\mathrm{N}$ lexicalization.

| relation | full links | outgoing |
|---|---|---|
| hyponymy | 26324 | 60062 |
| hypernymy | 26324 | 33725 |
| similarity | 10186 | 14785 |
| has category | 2131 | 2241 |
| category of | 2131 | 6135 |
| has instance | 641 | 5936 |
| instance of | 641 | 1131 |
| part meronymy | 2471 | 6029 |
| part holonymy | 2471 | 3408 |
| member meronymy | 400 | 734 |
| member holonymy | 400 | 1517 |
| subst. meronymy | 190 | 325 |
| subst. holonymy | 190 | 414 |

Lexical relations between particular terms cannot, in general, be transferred automatically, e.g. a region domain for a term in one language, signifying in what geographical region the term is used, will not apply to a second language. However, certain lexical relations such as the derivation relation still provide valuable information when interpreted as a general indicator of semantic relatedness, as can be seen in Table 5, which shows the results of a human evaluation for several different relation types. Incorrect relations are almost entirely due to incorrect term-sense mappings.

### 3.3 Structural Adequacy

As mentioned earlier, our machine learning approach is very parsimonious with respect to $L_\mathrm{N}$-specific prerequisites, and hence scales well to new languages. Some lexicographers contend that using one wordnet as the structural basis for another wordnet does not do justice to the structure of the new language's lexicon.

The most significant issue is certainly that the source wordnet may lack senses for certain terms in the new language, as in the case of the German "`Feierabend`". This point has already been addressed in Section 3.1.

Apart from this, it seems that general structural differences between languages rarely cause problems. When new wordnets are built independently from existing wordnets, many of the structural differences will not be due to actual

**Table 5.** Quality assessment for imported relations: For each relation type, 100 randomly selected links between two senses with $L_{\mathrm{N}}$-lexicalizations were evaluated.

| relation | accuracy |
|---|---|
| hyponymy, hypernymy | 84% |
| similarity | 90% |
| category | 91% |
| instance | 93% |
| part meronymy, holonymy | 83% |
| member meronymy, holonymy | 89% |
| subst. meronymy, holonymy | 83% |
| antonymy (as sense opposition) | 95% |
| derivation (as semantic similarity) | 96% |

conceptual differences between languages, but rather result from subjective decisions made by the individual human modellers [8].

Some of the rare examples of cultural differences affecting relations between two senses include perhaps the question of whether the local term for "`guinea pig`" should count as a hyponym of the respective term for "`pet`". For such cases, our suggestion is to manually add relation attributes that describe the idea of a connection being language-specific, culturally biased, or based on a specific taxonomy rather than holding unconditionally.

A more general issue is the adequacy of the four lexical categories (parts of speech) considered by Princeton WordNet. Fortunately, most of the differences between languages in this respect either concern functional words, or occur at very fine levels of distinctions, e.g. genus distinctions for German nouns, and thus are conventionally considered irrelevant to wordnets, though such information could be derived from monolingual dictionaries and added to the wordnet.

## 4 Human Consultation

One major disadvantage of automatically built wordnets is the lack of native-language glosses and example sentences, although this problem is not unique to automatically-built wordnets. Because of the great effort involved in compiling such information, manually built wordnets such as GermaNet also lack glosses and example sentences for the overwhelming majority of the senses listed. In this respect, automatically produced aligned wordnets have the advantage of at least making English-language glosses accessible.

Another significant issue is the quality of the mappings. As people are more familiar with high-quality print dictionaries, they do not expect to encounter incorrect entries when consulting a WordNet-like resource.

In contrast, we found that machine-generated wordnets can instead be used to provide machine-generated thesauri, where users expect to find more generally related terms rather than precise synonyms and gloss descriptions. In order to generate such a thesaurus, we relied on a simple technique that looks up all senses of a term as well as certain related senses, and then forms the union of all lexicalizations of these senses (Algorithm 4.1 with $n_h = 2$, $n_o = 2$, $n_g = 1$). Table 6 provides a sample entry from the German thesaurus resulting from our wordnet, and demonstrates that such resources can indeed be used for example as built-in thesauri in word processing applications.

**Algorithm 4.1** Thesaurus Generation

**Input:** a wordnet instance $W$ (with function $\sigma$ for retrieving senses and $\sigma^{-1}$ for retrieving the set of all terms for a sense), number of hypernym levels $n_h$, number of hyponym levels $n_o$, number of levels for other general relations $n_g$, set of acceptable general relations $R$
**Objective:** generate a thesaurus that lists related terms for any given term

**procedure** GenerateThesaurus($W, R$)
  **for each** term $t$ from $W$ **do** ▷ for every term $t$ listed in the wordnet
    $T \leftarrow \emptyset$ ▷ the list of related terms for $t$
    **for each** sense $s \in \sigma(t)$ **do** ▷ for each sense of $t$
      **for each** sense $s' \in$ Related($W, s, n_h, n_o, n_g, R$) **do**
        $T \leftarrow T \cup \sigma^{-1}(s')$ ▷ add lexicalizations of $s'$ to $T$
    output $T$ as list of related terms for $t$
**function** Related($W, s, n_h, n_o, n_g, R$)
  $S \leftarrow \{s\}$
  **for each** sense $s'$ related to $s$ with respect to $W$ **do** ▷ recursively visit related senses
    **if** ($s'$ hypernym of $s$) $\wedge$ ($n_h > 0$) **then**
      $S \leftarrow S \cup$ Related($W, s', n_h - 1, 0, 0, \emptyset$)
    **else if** ($s'$ hyponym of $s$) $\wedge$ ($n_o > 0$) **then**
      $S \leftarrow S \cup$ Related($W, s', 0, n_o - 1, 0, \emptyset$)
    **else if** $\exists r \in R$ : ($s'$ stands in relation $r$ to $s$) $\wedge$ ($n_g > 0$) **then**
      $S \leftarrow S \cup$ Related($W, s', 0, 0, n_g - 1, R$)
  **return** $S$

**Table 6.** Sample entries from generated thesaurus (which contains entries for 55522 terms, each entry listing 17 additional related terms on average)

| **headword:** Leseratte |
| --- |
| Buchgelehrte, Buchgelehrter, Bücherwurm, Geisteswissenschaftler, Gelehrte, Gelehrter, Stubengelehrte, Stubengelehrter, Student, Studentin, Wissenschaftler |

| **headword:** leserlich |
| --- |
| Lesbarkeit, Verständlichkeit |
| deutlich, entzifferbar, klar, lesbar, lesenswert, unlesbar, unleserlich, übersichtlich |

# 5 Monolingual Applications

## 5.1 General Remarks

Although at first it might seem that having wordnets aligned to the original WordNet is mainly beneficial for cross-lingual tasks, it turns out that the alignment also proves to be a major asset for monolingual applications, as one can leverage much of the information associated with the Princeton WordNet, e.g. the included English-language glosses, as well as a wide range of third-party resources, incl. topic domain information [15], links to ontologies such as SUMO [16] and YAGO [17], etc.

For instance, for the task of word sense disambiguation, a preliminary study using an algorithm that maximizes the overlap of the English-language glosses [18] showed promising results, although we were unable to evaluate it more adequately due to the lack of an appropriate sense-tagged test corpus. One problem we encountered, however, was that the generated wordnet sometimes did not cover all of the terms and senses to be disambiguated, which means that it is not an ideal sense inventory for word sense disambiguation tasks.

Apart from that, generated wordnets can be used for most other tasks that the English WordNet is usually employed for, including text and multimedia retrieval, text classification, text summarization, as well as semantic relatedness estimation, which we will now consider in more detail.

### 5.2 Semantic Relatedness

Several studies have attempted to devise means of automatically approximating semantic relatedness judgments made by humans, predicting e.g. that most humans consider the two terms "`fish`" and "`water`" semantically related. Such relatedness information is useful for a number of different tasks in information retrieval and text mining, and various techniques have been proposed, many relying on lexical resources such as WordNet. For the German language, Gurevych [19] reported that Lesk-style similarity measures based on the similarity of gloss descriptions [20] do not work well in their original form because GermaNet features only very few glosses, and those that do exist tend to be rather short. With machine-generated aligned wordnets, however, one can apply virtually any existing measure of relatedness that is based on the English WordNet, because English-language glosses and co-occurrence data are available.

We proceeded using the following assessment technique. Given two terms $t_1$, $t_2$, we estimate their semantic relatedness using the maximum relatedness score between any of their two senses:

$$\mathrm{rel}(t_1, t_2) = \max_{s_1 \in \sigma(t_1)} \max_{s_2 \in \sigma(t_2)} \mathrm{rel}(s_1, s_2) \tag{3}$$

For the relatedness scores, we consider three different approaches.

1. Graph distance: We consider the graph constituted by WordNet's senses and sense relations, and compute proximity scores for nodes in the graph by taking the maximum of the products of relation-specific edge weights for any two paths between two nodes.
2. Gloss Similarity: For each sense in WordNet, extended gloss descriptions are created by concatenating the glosses and lexicalizations associated with the sense as well as those associated with certain related senses (senses connected

via hyponymy, derivation/derived, member/part holonymy, and instance relations, as well as two levels of hypernyms). Each gloss description is then represented as a bag-of-words vector, where each dimension represents the TF-IDF value of a stemmed term from the glosses. For two senses $s_1$, $s_2$, one then computes the inner product of the two corresponding gloss vectors $\mathbf{c}_1$, $\mathbf{c}_2$ to determine the cosine of the angle $\theta_{\mathbf{c}_1,\mathbf{c}_2}$ between them, which characterizes the amount of term overlap for the two context strings:

$$\cos\ \theta_{\mathbf{c}_1,\mathbf{c}_2} = \frac{\langle \mathbf{c}_1, \mathbf{c}_2 \rangle}{||\mathbf{c}_1|| \cdot ||\mathbf{c}_2||} \tag{4}$$

3. Maximum: Since the two measures described above are based on very different information, we combined them into a meta-method that always chooses the maximum of these two relatedness scores.

For evaluating the approach, we employed three German datasets [19, 21] that capture the mean of relatedness assessments made by human judges. In each case, the assessments computed by our methods were compared with these means, and Pearson's sample correlation coefficient was computed. The results are displayed in Table 7, where we also list the current state-of-the-art scores obtained for GermaNet and Wikipedia as reported by Gurevych et al. [22].

**Table 7.** Evaluation of semantic relatedness measures, using Pearson's sample correlation coefficient in %. We compare our three semantic relatedness measures based on the automatically generated wordnet with the agreement between human annotators and scores for two alternative measures as reported by Gurevych et al. [22], one based on Wikipedia, the other on GermaNet.

| Dataset | GUR65 | | GUR350 | | ZG222 | |
|---|---|---|---|---|---|---|
| | Pearson $r$ | Coverage | Pearson $r$ | Coverage | Pearson $r$ | Coverage |
| Inter-Annot. Agreement | 0.81 | (65) | 0.69 | (350) | 0.49 | (222) |
| Wikipedia (ESA) | 0.56 | 65 | 0.52 | 333 | 0.32 | 205 |
| GermaNet (Lin) | 0.73 | 60 | 0.50 | 208 | 0.08 | 88 |
| Gen. wordnet (graph) | 0.72 | 54 | 0.64 | 185 | 0.41 | 89 |
| Gen. wordnet (gloss) | 0.77 | 54 | 0.59 | 185 | 0.47 | 89 |
| Gen. wordnet (max.) | 0.75 | 54 | 0.67 | 185 | 0.44 | 89 |

The results show that our semantic relatedness measures lead to near-optimal correlations with respect to the human inter-annotator agreement correlations. The main drawback of our approach is a reduced coverage compared to Wikipedia and GermaNet, because scores can only be computed when both parts of a term pair are covered by the generated wordnet.

One advantage of our approach is that it may also be applied without any further changes to the task of cross-lingually assessing the relatedness of English terms with German terms. In the following section, we will take a closer look at the general suitability of our wordnet for multilingual applications.

## 6 Multilingual Applications

### 6.1 General Remarks

We can distinguish the following two categories of applications with multilingual support.

- multilingual applications that need to support certain operations on more than just a single language, e.g. word processors with thesauri for multiple languages
- multilingual applications that perform cross-lingual operations

By creating isolated wordnets for many different languages one addresses only the first case. For the second case, one can use multiple wordnets for different languages where the senses are strongly interlinked. The ideal case is when there is no sense duplication, i.e. if two words in different languages share the same meaning, they should be linked to the same sense. The techniques described in Section 2 achieve this by producing wordnets that are strictly aligned to the source wordnet whenever appropriate.

Aligned wordnets thus can be used for various cross-lingual tasks, including cross-lingual information retrieval [23], and cross-lingual text classification, which will now be studied.

### 6.2 Cross-Lingual Text Classification

Text classification is the task of assigning text documents to the classes or categories considered most appropriate, thereby e.g. topically distinguishing texts about thermodynamics from others dealing with quantum mechanics. This is commonly achieved by representing each document using a vector in a high-dimensional feature space where each feature accounts for the occurrence of a particular term from the document set (a bag-of-words model), and then applying machine learning techniques such as support vector machines. For more information, please refer to Sebastiani's survey [24].

Cross-lingual text classification is a much more challenging task. Since documents from two different languages obviously have completely different term distributions, the conventional bag-of-words representations perform poorly. Instead, it is necessary to induce representations that tend to give two documents from different languages similar representations when their content is similar.

One means of achieving this is the use of language-independent conceptual feature spaces where the feature dimensions represent meanings of terms rather than just the original terms. We process a document by removing stop words, performing part-of-speech tagging and lemmatization using the TreeTagger [25], and then map each term to the respective sense entries listed by the wordnet instance. In order to avoid decreasing recall levels, we do not disambiguate in any way other than acknowledging the lexical category of a term, but rather assign each sense $s$ a local score $\frac{w_{t,s}}{\sum\limits_{s' \in \sigma(t)} w_{t,s'}}$ whenever a term $t$ is mapped to multiple senses $s \in \sigma(t)$. Here, $w_{t,s}$ is the weight of the link from $t$ to $s$ as provided by the wordnet if the lexical category between document term and sense match, or 0 otherwise. We test two different setups: one relying on regular unweighted wordnets ($w_{t,s} \in \{0, 1\}$), and another based on a weighted German wordnet ($w_{t,s} \in [0, 1]$), as described in Section 2. Since the original document terms may include useful language-neutral terms such as names of people or organizations, they are also taken into account as tokens with a weight of 1. By summing up the weights for each local occurrence of a token $t$ (a term or a sense) within a document $d$, one arrives at document-level token occurrence scores $n(t, d)$, from which one can then compute TF-IDF-like feature vectors using the following formula:

$$\log(n(t,d)+1)\ \log\left(\frac{|D|}{|\{d \in D \mid n(t,d) \geq 1\}|}\right) \tag{5}$$

where $D$ is the set of training documents.

This approach was tested using a cross-lingual dataset derived from the Reuters RCV1 and RCV2 collections of newswire articles [26, 27]. We randomly selected 15 topics shared by the two corpora in order to arrive at $\binom{15}{2} = 105$ binary classification tasks, each based on 200 training documents in one language, and 600 test documents in a second language, likewise randomly selected, however ensuring equal numbers of positive and negative examples in order to avoid biased error rates. We considered a) German training documents and English test documents and b) English training documents and German test documents. For training, we relied on the SVMlight implementation [28] of support vector machine learning [12], which is known to work very well for text classification.

**Table 8.** Evaluation of cross-lingual text classification in terms of micro-averaged accuracy, precision, recall, and $F_1$-score for a German-English as well as an English-German setup. We compare the standard bag-of-words TF-IDF representation with two wordnet-based representations, one using an unweighted, the other based on a weighted German wordnet.

| | acc. | prec. | rec. | $F_1$ |
|---|---|---|---|---|
| *German-English* | | | | |
| TF-IDF | 80.56 | 77.49 | 86.14 | 81.59 |
| Wordnet (unweighted) | 87.09 | 85.27 | 89.68 | 87.42 |
| Wordnet (weighted) | 87.98 | 85.48 | 91.51 | 88.39 |
| *English-German* | | | | |
| TF-IDF | 78.82 | 79.19 | 78.20 | 78.69 |
| Wordnet (unweighted) | 85.39 | 87.38 | 82.74 | 84.99 |
| Wordnet (weighted) | 87.47 | 87.73 | 87.07 | 87.40 |

The results in Table 8 clearly show that automatically built wordnets aid in cross-lingual text classification. Since many of the Reuters topic categories are business-related, using only the original document terms, which include names of companies and people, already works surprisingly well, though presumably not well enough for use in production settings. By considering wordnet senses, both precision and recall are boosted significantly. This implies that English terms in the training set are being mapped to the same senses as the corresponding German terms in the test documents. Using the weighted wordnet version further improves the recall, as more relevant terms and senses are covered.

## 7 Conclusions

We have shown that machine-generated wordnets are useful for a number of different purposes. First of all, of course, they can serve as a valuable starting point for establishing more reliable wordnets, which would involve manually extending the coverage and addressing issues arising from differences between the lexicons of different languages.

At the same time, machine-generated wordnets can be used directly without further manual work to generate thesauri for human use, or for a number of different natural language processing applications, as we have shown in particular for semantic relatedness estimation and cross-lingual text classification.

In the future, we would like to investigate techniques for extending the coverage of such statistically generated wordnets to senses not covered by the original Princeton WordNet. We hope that our research will aid in contributing to making lexical resources available for languages which to date have not been dealt with by the wordnet community.

## References


1. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
2. Vossen, P.: Right or wrong: Combining lexical resources in the EuroWordNet project. In: Proc. Euralex-96. (1996) 715–728
3. Okumura, A., Hovy, E.: Building Japanese-English dictionary based on ontology for machine translation. In: Proc. Workshop on Human Language Technology, HLT, Morristown, NJ, USA, Association for Computational Linguistics (1994) 141–146
4. Rigau, G., Agirre, E.: Disambiguating bilingual nominal entries against WordNet. In: Proc. Workshop on the Computational Lexicon at the 7th European Summer School in Logic, Language and Information, ESSLLI. (1995)
5. Atserias, J., Climent, S., Farreres, X., Rigau, G., Rodríguez, H.: Combining multiple methods for the automatic construction of multilingual WordNets. In: Proc. International Conference on Recent Advances in NLP. (1997) 143–149
6. Benitez, L., Cervell, S., Escudero, G., Lopez, M., Rigau, G., Taulé, M.: Methods and tools for building the Catalan WordNet. In: Proc. ELRA Workshop on Language Resources for European Minority Languages at LREC 1998. (1998)
7. Miháltz, M., Prószéky, G.: Results and evaluation of Hungarian Nominal WordNet v1.0. In: Proc. Second Global WordNet Conference, Brno, Czech Republic, Masaryk University (2004)
8. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: Developing an aligned multilingual database. In: Proc. First International Global WordNet Conference, Mysore, India. (2002) 293–302

9. Ordan, N., Wintner, S.: Hebrew WordNet: a test case of aligning lexical databases across languages. International Journal of Translation **19**(1) (2007)
10. de Melo, G., Weikum, G.: A machine learning approach to building aligned wordnets. In: Proc. International Conference on Global Interoperability for Language Resources, ICGL. (2008)
11. Richter, F.: Ding Version 1.5, `http://www-user.tu-chemnitz.de/~fri/ding/`. (2007)
12. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
13. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001)
14. Hamp, B., Feldweg, H.: GermaNet — a lexical-semantic net for German. In: Proc. ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid (1997)
15. Bentivogli, L., Forner, P., Magnini, B., Pianta, E.: Revising the Wordnet Domains hierarchy: semantics, coverage and balancing. In: Proc. COLING 2004 Workshop on Multilingual Linguistic Resources, Geneva, Switzerland (2004) 94–101
16. Niles, I., Pease, A.: Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: Proc. 2003 International Conference on Information and Knowledge Engineering (IKE '03), Las Vegas, NV, USA. (2003)
17. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: 16th International World Wide Web conference, WWW, New York, NY, USA, ACM Press (2007)
18. Patwardhan, S., Banerjee, S., Pedersen, T.: Using measures of semantic relatedness for word sense disambiguation. In: Proc. 4th Intl. Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Mexico City, Mexico. (2003)
19. Gurevych, I.: Using the structure of a conceptual network in computing semantic relatedness. In: Proc. Second International Joint Conference on Natural Language Processing, IJCNLP, Jeju Island, Republic of Korea (2005)
20. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proc. 5th annual international conference on Systems documentation, SIGDOC '86, New York, NY, USA, ACM Press (1986) 24–26
21. Zesch, T., Gurevych, I.: Automatically creating datasets for measures of semantic relatedness. In: COLING/ACL 2006 Workshop on Linguistic Distances, Sydney, Australia (2006) 16–24
22. Gurevych, I., Müller, C., Zesch, T.: What to be? - Electronic career guidance based on semantic relatedness. In: Proc. 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, Association for Computational Linguistics (2007) 1032–1039
23. Chen, H.H., Lin, C.C., Lin, W.C.: Construction of a Chinese-English WordNet and its application to CLIR. In: Proc. Fifth International Workshop on Information Retrieval with Asian languages, IRAL '00, New York, NY, USA, ACM Press (2000) 189–196
24. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34**(1) (2002) 1–47
25. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Intl. Conference on New Methods in Language Processing, Manchester, UK (1994)
26. Reuters: Reuters Corpus, vol. 1: English Language, 1996-08-20 to 1997-08-19 (2000)
27. Reuters: Reuters Corpus, vol. 2: Multilingual, 1996-08-20 to 1997-08-19 (2000)
28. Joachims, T.: Making large-scale support vector machine learning practical. In Schölkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge, MA, USA (1999)