Lang Resources & Evaluation (2012) 46:287–311
DOI 10.1007/s10579-012-9183-2

ORIGINAL PAPER

Constructing and utilizing wordnets using statistical methods

Gerard de Melo · Gerhard Weikum

Published online: 10 May 2012

Abstract Lexical databases following the wordnet paradigm capture information about words, word senses, and their relationships. A large number of existing tools and datasets are based on the original WordNet, so extending the landscape of resources aligned with WordNet leads to great potential for interoperability and to substantial synergies. Wordnets are being compiled for a considerable number of languages, however most have yet to reach a comparable level of coverage. We propose a method for automatically producing such resources for new languages based on WordNet, and analyse the implications of this approach both from a linguistic perspective as well as by considering natural language processing tasks. Our approach takes advantage of the original WordNet in conjunction with translation dictionaries. A small set of training associations is used to learn a statistical model for predicting associations between terms and senses. The associations are represented using a variety of scores that take into account structural properties as well as semantic relatedness and corpus frequency information. Although the resulting wordnets are imperfect in terms of their quality and coverage of language-specific phenomena, we show that they constitute a cheap and suitable alternative for many applications, both for monolingual tasks as well as for cross-lingual interoperability. Apart from analysing the resources directly, we conducted tests on semantic relatedness assessment and cross-lingual text classification with very promising results.

Keywords Lexical resources · WordNet · Machine learning

G. de Melo (✉) · G. Weikum
Max Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany
e-mail: demelo@mpi-inf.mpg.de

G. Weikum
e-mail: weikum@mpi-inf.mpg.de

1 Introduction

Lexical databases are indispensable for many natural language processing tasks. WordNet (Fellbaum 1998) is the most well-known and most widely used lexical database for English language processing, and is the fruit of over 20 years of manual work carried out at Princeton University. A large number of existing tools and datasets are based on WordNet, so extending the landscape of resources aligned with WordNet leads to great potential for interoperability and to substantial synergies. The original WordNet for the English language inspired endeavours to create similarly structured resources ("wordnets") for other languages, e.g. in the context of the EuroWordNet EU project (Vossen 1998), the BalkaNet project (Tufiş et al. 2004), as well as under the auspices of the Global WordNet Association. Nevertheless, we contend that despite several decades of work on such resources, there is still a great need for additional research into more efficient means of producing them. Consider, for instance, that there are about 7,000 living languages, but only around 50 for which wordnet versions have been created, many indeed still in a preliminary stage with very low coverage, and only about a handful of languages with wordnet versions that are freely downloadable from the Internet. Furthermore, several existing wordnets unfortunately form completely independent networks that are not connected to and hence not interoperable with other wordnets.

The main bottleneck is the laborious compilation process, which requires skilled experts to work on such a resource for several years. In order to complement the existing manually compiled wordnets, we thus propose a new approach to constructing wordnets that trades off accuracy for a much faster compilation process, and hence frequently leads to more terms being covered than in existing wordnets. Our approach is based on learning classifications, and therefore is completely automatic once an initial set of training associations is provided. The fact that the wordnets are aligned with the original Princeton WordNet greatly facilitates interoperability with existing wordnets (e.g. English-language glosses are available) as well as many additional resources such as ontologies and mappings, as detailed in Sect. 2.

Certainly, the resulting wordnets will not have the same level of accuracy as resources carefully constructed by skilled lexicographers, however they can (1) serve as a valuable starting point for creating more accurate ones, and (2) be used immediately in many natural language processing tasks where coverage is more important than perfect accuracy, as will later be demonstrated in Sect. 6.

The remainder of this article is organized as follows. Section 2 begins with a brief introduction to wordnets and their role for interoperability. After a brief summary of alternative compilation techniques in Sect. 3, the main focus of this article will be a thorough description of an automatic statistical approach to constructing wordnets in Sect. 4. The implications of using such an approach as well as evaluation results are studied in great detail in Sect. 5. Section 6 considers possible applications of automatically built wordnets, discussing human use as well as experimental results on natural language processing tasks such as semantic relatedness and cross-lingual text classification. Finally, concluding remarks are provided in Sect. 7.

2 Wordnets and their role for interoperability

We will begin by introducing Princeton WordNet, the original wordnet that inspired all successors, as well as by discussing the role of wordnets for interoperability.

2.1 Princeton WordNet

Princeton WordNet (Fellbaum 1998) is a lexical database for the English language that captures information about how words and word senses in the English language are related. It lists the senses that a word can assume and identifies senses that are synonymous in meaning as semantic units called *synsets*. Terms and synsets are organized as a network of nodes linked by various lexico-semantic relations.

The *hyponymy* relation can be defined as one that "holds between a more specific, or subordinate, lexeme and a more general, or superordinate, lexeme, as exemplified by such pairs as 'cow':'animal', 'rose':'flower"' (Lyons 1977). *Hypernymy* is the respective inverse relation. In WordNet, these are captured as relations between word senses. The antonymy relation represents semantic opposition between terms. Other relations include instance relationships and several kinds of meronymic relations.

2.2 Wordnets and interoperability

There is a significant amount of ongoing work on standards that will facilitate interoperability for language resources and natural processing applications. Apart from agreeing on common data formats, an important challenge is the establishment of shared identifiers that allow us to unambiguously refer to linguistic phenomena. Examples include the ISO 639 standards for language codes and the development of the ISO Data Category Registry to provide labels for parts of speech, syntactic constituency, etc. (Francopoulo et al. 2008).

At the same time, there is also an increasing need to refer to word senses in an unambiguous way, e.g. in translation resources. We believe that WordNet qualifies as a suitable starting point for developing a multilingual sense inventory. Wordnets in several languages are already connected to the original one. Geographical information (Buscaldi and Rosso 2008) and pictures (Deng et al. 2009) are available for many sense identifiers listed in WordNet. Other resources linked to WordNet include topical domain labels (Bentivogli et al. 2004), verb lexicons such as VerbNet (Kipper et al. 2000) and FrameNet (Baker and Fellbaum 2008), and ontologies like SUMO (Niles and Pease 2003), YAGO (Suchanek et al. 2007), DOLCE (Gangemi et al. 2003), and OpenCyc (Cycorp Inc. 2008). Via YAGO, WordNet is also connected to Wikipedia and many other datasets in the Linked Data Web (Bizer et al. 2009).

By building new wordnets that are aligned with the English WordNet, we can not only contribute to this infrastructure and increase its value, but also benefit from it when deploying the new wordnets for natural language processing.

3 Previous work on building wordnets automatically

Prior to introducing our statistical approach to constructing wordnets, we will summarize some of the previous means of creating wordnets.

One general strategy is the so-called *merge model*, where an existing thesaurus is converted to a wordnet-like format and then semi-automatically linked to other wordnets or to an interlingual synset index. The downside of this strategy is that it cannot be applied to a large range of languages, unless some pre-existing wordnet-like thesaurus for each of these languages is found or established.

An alternative general strategy is the *expand model*, which requires much fewer pre-existing resources. 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 the terms associated with s from L_0 to a new language L_N using a translation dictionary. (3) Additionally retain all appropriate semantic relations between senses from the existing wordnet in order to arrive at a new wordnet for L_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_N-term t does not imply that t applies to all senses of e. For example, with regard to the translation from the English word "*bank*" to the German "*Bank*", we 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. Knight (1993) created an ontology for machine translation by linking entries in Longman's Dictionary of Contemporary English to WordNet, taking into account gloss definitions as well as the semantic hierarchy information present in the dictionary, though unfortunately not available in the settings we consider (cf. Sect. 4.2). Okumura and Hovy (1994) used a Japanese-English dictionary to link a Japanese lexicon to this ontology, based on several heuristics, most importantly monosemy, i.e. considering when the ontology lists only one candidate concept for an English translation, and equivalent word matches, i.e. accepting the concepts shared by multiple translations of a word.

Another important line of research starting with Rigau and Agirre (1995), and extended by Atserias et al. (1997) resulted in automatic techniques for creating preliminary noun-only versions of the Spanish WordNet and later also the Catalan WordNet (Benitez et al. 1998). 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_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 (Miháltz and Prószéky 2004).

Pianta et al. (2002) used similar ideas in conjunction with a cosine similarity-based heuristic to produce rankings of the most likely candidate senses. 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 methodology was used to create MultiWordNet Italian and later also adopted for the Hebrew WordNet (Ordan and Wintner 2007).

Sathapornrungkij and Pluempitiwiriyawej (2005) used criteria proposed by Atserias et al. (1997), and then performed a regression analysis in order to reduce the number of accepted associations and thus increase the accuracy. Since they merely relied on 12 binary criteria rather than numeric scores, they were unable to obtain a higher recall by applying their model to other term-sense pairs not fulfilling one of the chosen criteria.

A more advanced approach that requires only minimal human work lies in using machine learning algorithms based on a large number of scores to identify more subtle decision rules. These decision rules can rely on a number of different heuristic scores with different thresholds.

4 Building wordnets by learning classifications

4.1 General outline

In order to build wordnets automatically, we suggest the following approach. Let L_N denote the language for which a wordnet is to be constructed, and L_0 denote the language of an existing wordnet that serves as a template for the new one, in our case the English language due to our choice of Princeton WordNet as the template. Acknowledging the caveats pointed out in Sect. 5, we can treat this existing wordnet as providing an inventory of possible senses.

The most important desideratum obviously are the links from terms in L_N to their respective senses. This challenge is tackled by means of translation dictionaries, which we use to obtain translations of terms from L_N to terms from L_0. These translations in turn allow us to construct for each of the original L_N-terms a *candidate set* of synsets that are potentially valid senses.

The central difficulty then is determining which of the candidate synsets to accept and which not. Given the polysemy of terms in L_0, it often turns out that the majority of the candidate synsets are not acceptable as senses for the L_N-term. Our approach relies on a set of training associations between L_N-terms and synsets to learn a disambiguation model that can then provide confidence scores indicating how certain we can be about a particular association being correct.

To create this disambiguation model, we compute several numeric scores (*feature values*) for a given association between an L_N term t and a candidate synset s, which together constitute a *feature vector*. Based on a small set of manually established labels for such (t, s)-pairs, we create the corresponding training set of feature vectors. The disambiguation model can then be derived using well-known classification learning techniques that consider statistical properties of the training vectors. Such a model can be used to make predictions for any other (t, s)-pair. To create the new wordnet, the model is applied to all pairs (t, s) consisting of an L_N term t and one of its candidate synsets s. In a final step, one can then import certain relations between synsets from the existing wordnet.

This approach has several advantages compared to the previous work in this field (cf. Sect. 3). First of all, the previous automatic approaches were based on hard acceptance criteria—either a (t, s)-pair satisfies a criterion or not. Many attributes of word senses do not lend themselves easily to such an antagonistic view, e.g. sense relatedness measures produce numeric scores, and thus can be better accommodated in a model that uses real-valued feature vectors. Furthermore, while Atserias et al. (1997) investigate combinations of two heuristics to arrive at a greater accuracy, a classification learning approach can take into account suitable combinations of even more heuristics, indeed arbitrary linear (or even non-linear) combinations of feature values.

Following this general description of the overall procedure, the following sections will expound on several aspects in much greater detail.

4.2 Candidate sets

Given a translation from a term t from L_N to a term e from L_0, it is safe to assume that there is some semantic overlap between t and e, and hence there is a reasonably high probability that some sense of e is also a sense of t.

Our approach makes use of translation dictionaries, however with the constraint of relying on a minimal amount of information specific to L_N so that the procedure remains generalizable to as many languages as possible. The dictionary is thus conceived as offering a simple n:m-mapping between terms in L_0 and terms in L_N, with optional part of speech information, as in the following German-English excerpt:

{n}	Schulabbrecher	-	dropout
	...		...
{n}	Schulklasse	-	class
{n}	Schulklasse	-	form
	...		...
	schulmäßig	-	scholastic
{adv}	schulmäßig	-	scholastically

We thus proceed as follows: for each term t from L_N, retrieve the set of translations $\phi(t)$. For each L_0-translation e in such a $\phi(t)$, retrieve the set of senses $\sigma(e)$ from our existing wordnet, e.g. for the German term "*Schulklasse*" the senses of the translations "*class*" and "*form*" would be considered.

The union $\bigcup_{e\in\phi(t)} \sigma(e)$ then constitutes the candidate set $C(t)$ for a particular term t, and our goal will be to determine for each sense $s \in C(t)$ whether it is appropriate to consider s a sense of t. This is undoubtedly a very difficult task, as the dictionaries provide only limited information that could aid in determining which of the often many different senses apply, e.g. WordNet lists 9 senses for the word "*class*" and 23 senses for "*form*".

4.3 Feature computation

In our approach, this task of determining the appropriate senses among the candidates is construed as a binary classification problem. A real-valued feature vector **x** is created for each pair (t, s) of a term t from L_{N} and a relevant candidate sense $s \in C(t)$. For example, if t represents "*Schulklasse*", then s could be one of the senses of "*class*". In order to create the feature vectors, a variety of different scores x_i are used as features and combined as components of numeric vectors $\mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$. These scores x_i are intended to quantify some information about the respective term-sense pair.

4.3.1 Sense weighting functions

Several features that will be described later on depend on some kind of assessment of the importance of senses s with respect to the particular L_{N}-term t under consideration. We consider the following weighting functions $\gamma(t, s)$:

- $\gamma_1(t, s) = 1$ is used for unweighted features
- $\gamma_{\mathrm{lc}}(t, s)$ represents an estimation of the lexical category compatibility between t and s as a value in [0, 1], where 0 means they are incompatible, e.g. when t is a noun and s is an adjective sense, and 1 means they are fully compatible (see Sect. 4.3.6 for more information on how these values are obtained).
- $\gamma_{\mathrm{r}}(t, s)$ considers the ranks of the senses as listed by WordNet for the translations of t, as these are indicators for the importance of a sense. It is computed as follows:

$$\gamma_{\mathrm{r}}(t, s) = \gamma_{\mathrm{lc}}(t, s)\left[\sum_{e \in \phi(t)} \frac{1}{r(e, s)}\right]$$

  where $r(e, s)$ yields 1 if s is the highest-ranked sense for e, 2 for the second sense, and so on.
- $\gamma_{\mathrm{f}}(t, s)$ considers the corpus frequency information provided with WordNet:

$$\gamma_{\mathrm{f}}(t, s) = \gamma_{\mathrm{lc}}(t, s)\left[\sum_{e \in \phi(t)} \frac{f(e, s)}{\sum_{s' \in \sigma(e)} \lambda_{s,s'} f(e, s')}\right]$$

  where $f(e, s)$ returns the number of occurrences of term e with sense s in the SemCor corpus, and $\lambda_{s,s'}$ is 1 if the lexical categories of s and s' match, and 0 otherwise.

4.3.2 Semantic relatedness measures

Apart from weighting functions, our approach is fundamentally based on measures of semantic relatedness between senses, e.g. the single sense of "*schoolhouse*" is related to the educational institution sense of "*school*", but not to the sense of "*school*" that refers to groups of fish. Before going into details of how semantic relatedness contributes to many of our fitness scores, we shall first introduce several relatedness estimation heuristics.

- $\mathrm{sim}_{\mathrm{id}}(s_1, s_2)$ is simply the trivial identity indicator function, i.e. yields 1 if $s_1 = s_2$, and 0 otherwise.

$$\mathrm{sim}_{\mathrm{id}}(s_1, s_2) = \begin{cases} 1 & s_1 = s_2 \\ 0 & \text{otherwise} \end{cases}$$

- $\mathrm{sim}_{\mathrm{f}}(s_1, s_2)$ considers not only whether two senses are identical but also takes into account senses that stand in a parent-child or sibling relationship in terms of the hypernym hierarchy.

$$\mathrm{sim}_{\mathrm{f}}(s_1, s_2) = \begin{cases} 1 & s_1 = s_2 \\ 0.8 & \text{hypernymy/hyponymy} \\ 0.7 & \text{siblings, no hypernymy} \\ 0 & \text{otherwise} \end{cases}$$

- $\mathrm{sim}_{\mathrm{n}}(s_1, s_2)$ considers the neighbourhood in the graph constituted by WordNet's senses and sense relations. It acknowledges relations other than hypernymy/hyponymy as well as transitive connections (e.g. a holonym of a hypernym). For a given path in the graph, one can compute a proximity score multiplicatively from relation-specific edge weights (e.g. 0.8 for immediate hypernymy, 0.7 for immediate holonymy). The relatedness score is defined as the maximum proximity score for any path between s_1 and s_2 if this maximum is above or equal to a pre-defined threshold $\alpha_n = 0.35$, and 0 otherwise. It can be obtained efficiently using a Dijkstra-like algorithm (de Melo and Siersdorfer 2007).
- $\mathrm{sim}_{\mathrm{c}}(s_1, s_2)$ uses the cosine similarity of extended gloss context strings for senses. For each of the two senses s_1 and s_2, extended gloss descriptions are created by concatenating the WordNet glosses and lexicalizations associated directly with the senses 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). The terms in these glosses are stemmed using Porter's stemmer, and the two extended gloss descriptions are then recast as bag-of-words vectors $\mathbf{v}_1$, $\mathbf{v}_2$, where each dimension represents the TF-IDF score of a stemmed term from the extended glosses. One then computes the inner product of these two gloss vectors to determine the cosine of the angle $\theta_{\mathbf{v}1,\mathbf{v}2}$ between them, as it characterizes the amount of term overlap between the two context strings:

$$\mathrm{sim}_{\mathrm{c}}(s_1, s_2) = \cos\, \theta_{\mathbf{v}_1,\mathbf{v}_2} = \frac{\langle \mathbf{v}_1, \mathbf{v}_2 \rangle}{||\mathbf{v}_1|| \cdot ||\mathbf{v}_2||} \tag{1}$$

- $\mathrm{sim}_{\mathrm{m}}(s_1, s_2)$, finally, is a meta-measure that is simply defined as

$$\mathrm{sim}_{\mathrm{m}}(s_1, s_2) = \max\{\mathrm{sim}_{\mathrm{f}}(s_1, s_2), \mathrm{sim}_{\mathrm{n}}(s_1, s_2), \mathrm{sim}_{\mathrm{c}}(s_1, s_2)\} \tag{2}$$

and hence combines the power of $\mathrm{sim}_{\mathrm{f}}$, $\mathrm{sim}_{\mathrm{n}}$, and $\mathrm{sim}_{\mathrm{c}}$. It is particularly valuable due to the fact that $\mathrm{sim}_{\mathrm{n}}$ and $\mathrm{sim}_{\mathrm{c}}$ are based on very different characteristics of the senses.

4.3.3 *Semantic overlap features*

One important way of making use of the semantic relatedness measures is to exploit that an association should more likely be accepted when a term t has multiple English translations e, and the candidate sense s under consideration is somewhat pertinent to multiple of them. For instance, the German "*Schulklasse*" has the terms "*class*" and "*form*" as translations. While "*form*" can not only refer to a body of students who are taught together but also e.g. to a tax form, only the former of these two senses overlaps semantically with the senses of "*class*".

Given a term t and a candidate sense s, we integrate scores of the following form into the respective feature vector:

$$\sum_{e \in \phi(t)} \max_{s' \in \sigma(e)} \gamma(t, s')\, \mathrm{sim}(s, s') \tag{3}$$

$$\sum_{e \in \phi(t)} \frac{\sum_{s' \in \sigma(e)} \gamma(t, s')\, \mathrm{sim}(s, s')}{\sum_{s' \in \sigma(e)} \gamma(t, s')} \tag{4}$$

where $\mathrm{sim}(s_1, s_2)$ represents a semantic relatedness measure and the $\gamma(t, s)$ function provides weights as described earlier. The simple identity relatedness function $\mathrm{sim}_{\mathrm{id}}$ and the constant weighting function $\gamma_1(t, s) = 1$ make Eq. 3 yield a simple count of how many English terms are mapped to the sense, reminiscent e.g. of the equivalent word matching of Okumura and Hovy (1994) (cf. Sect. 3). By using the above formulae to produce a large number of feature values with all combinations of weighting functions and relatedness measures mentioned in Sects. 4.3.1 and 4.3.2, we are additionally able to account for cases where the terms are related but do not share senses.

4.3.4 *Polysemy-based scores*

Another set of features are based on the polysemy of the L_0-translations, i.e. on the idea that an association is more likely correct whenever there are few alternative senses to choose from. Akin to the monosemy heuristic of Okumura et al. (see Sect. 3), we can consider for instance the German "*Schulleiter*" with its translation "*headmaster*", which in turn only has one single sense listed in WordNet, so it is rather safe to accept this sense also for the German term. More generally, given a term t and a sense s, several scores can be computed as

$$\left(1 + \sum_{s' \in C(t)} \gamma(t, s')(1 - \mathrm{sim}(s, s'))\right)^{-1} \tag{5}$$

where $\gamma(t, s)$ is a weighting function and $C(t)$ stands for the complete candidate set. Another set of scores is computed as

$$\sum_{e\in\phi(t)} \frac{\mathbf{1}_{\sigma(e)}(s)}{1+\sum_{s'\in\sigma(e)}\gamma(t,s')(1-\mathrm{sim}(s,s'))} \tag{6}$$

where $\mathbf{1}_{\sigma(e)}(s)$ is the indicator function for $\sigma(e)$, and therefore yields 1 if $s \in \sigma(e)$ and 0 otherwise.

Again, we can use $\mathrm{sim}_{\mathrm{id}}(s_1, s_2)$ and $\gamma_1(t, s)$ to illustrate the simplest case: With these choices, Eq. 5 yields the reciprocal of the total number of candidate senses and in Eq. 6 the denominator of each addend becomes 1 whenever the respective term e is monosemous according to WordNet. More advanced scores are computed by

- using Eqs. 5, 6 with $\gamma_1(t, s)$, combined with either $\mathrm{sim}_{\mathrm{f}}$, $\mathrm{sim}_{\mathrm{c}}$, $\mathrm{sim}_{\mathrm{n}}$, or $\mathrm{sim}_{\mathrm{m}}$, and
- using Eq. 6 with $\mathrm{sim}_{\mathrm{id}}(s_1, s_2)$ and one of the weighting functions $\gamma_{\mathrm{lc}}(t, s)$, $\gamma_{\mathrm{r}}(t, s)$, or $\gamma_{\mathrm{f}}(t, s)$.

4.3.5 Additional features

We further consider a series of other, less essential features, including the following:

- scores based on the number of translations

$$\left(\sum_{e\in\phi(t)} \lambda(t,e)\right)^{-1}$$

  where $\lambda(t, e)$ is a translation weighting function that can be either $\lambda_{\mathrm{id}}(t, e) = 1$ or $\lambda_{\mathrm{wn}}(t, e)$, which is 1 if $\sigma(e) \neq \emptyset$, and 0 otherwise.
- the ratio

$$\frac{\sum_{e\in\phi(t)}\lambda_{\mathrm{wn}}(t,e)}{\sum_{e\in\phi(t)}\lambda_{\mathrm{id}}(t,e)} = \frac{\sum_{e\in\phi(t)}\lambda_{\mathrm{wn}}(t,e)}{|\phi(t)|}$$

  for the above definitions of λ_{wn} and λ_{id}.
- a score based on back-translations

$$\sum_{e\in\phi(t)} \frac{\mathbf{1}_{\sigma(e)}(s)}{|\phi^{-1}(e)|}$$

  where $\phi^{-1}(e)$ is defined as $\{t \,|\, e \in \phi(t)\}$.
- the number of lexicalizations of the candidate sense, i.e. $|\sigma^{-1}(s)|$, where $\sigma^{-1}(s)$ is defined as $\{e \,|\, s \in \sigma(e)\}$.
- the ratio of sense lexicalizations that are translations of t, i.e.

$$\frac{\sum_{e\in\sigma^{-1}(s)}\lambda_{\mathrm{tr}}(t,e)}{|\sigma^{-1}(s)|}$$

  where $\sigma^{-1}(s)$ is defined as above, and $\lambda_{\mathrm{tr}}(t, e)$ yields 1 if $e \in \phi(t)$ and 0 otherwise.

- indicator values that express whether the candidate sense s is a noun, verb, adjective, or adverb sense, respectively.

4.3.6 Lexical category compatibility

Unlike previous work, our study considers all lexical categories (parts of speech) covered by the existing wordnet rather than just nouns. This immediately leads to the problem that the number of candidate senses greatly increases, and we need to come up with some means of preventing a noun from being mapped to a verb sense in WordNet, for instance.

Our solution rests on two pillars. Obviously, whenever the translation dictionary explicitly provides lexical category information, one can simply use hard-coded compatibility indicators, e.g. we give any German adjective a compatibility value of 0.0 with English noun senses, but 1.0 with English adjective as well as adverb senses.

In light of the fact that such explicit information may not always be available, we resort to additional heuristics when necessary, thereby ensuring that our approach remains applicable to a broad range of different scenarios. For each lexical category, a C4.5 decision tree is used to estimate the compatibility based on superficial attributes of the terms such as suffixes and capitalization. In many languages, such attributes provide hints about the part of speech of a word. Growing the trees does not require any manually created training data, because we can leverage terms where all candidate senses share the same lexical category as examples. The features employed are given in the following list. Note that since the terms in L_{N} can be multi-word expressions, much of this information is captured separately for the first and last word of any candidate expression.

- prefixes of the first and last word up to a length of 10, e.g. for the German verb "*einschulen*", "*e*", "*ei*", "*ein*", etc. would be considered
- suffixes of the first and last word up to a length of 10 (without case conversion), e.g. "*n*", "*en*", "*len*", etc. for "*einschulen*".
- capitalization of the first and last word (Boolean features for no capitalization, capitalized first character, and all characters capitalized)
- term length

The decision trees were pruned to have confidence levels of at least 0.25 with at least 2 instances per leaf. The confidence estimations from the leaves can then be used to determine a lexical category compatibility score as a feature in the feature vector. For languages where the predictions are too unreliable, we may instead use a constant value of 0.5.

4.4 Learning the disambiguation model

Having defined a feature computation procedure, we can apply well-known classification learning techniques to derive the disambiguation model.

A classification is an assignment of class labels $y \in \mathcal{Y}$ to objects $x \in \mathcal{X}$, and can be formalized as a function $\widehat{f} : \mathcal{X} \times \mathcal{Y} \longrightarrow [0, 1]$ that, given such an x and y, yields a value that provides the degree of confidence in the assignment being correct. We consider only binary problems, where $\mathcal{Y} = \{A, \overline{A}\}$ for some class A and its complement $\overline{A}$, and only consider the single label case, where each object is assigned exactly one class. Learning a classification then consists in finding a function f that approximates a true classification $\widehat{f}$ with low approximation error, given a set of correctly labelled training examples $(x, y) \in \mathcal{X} \times \mathcal{Y}$. In our case, the objects are term-sense pairs $x = (t, s)$, and the class y is either A or its complement $\overline{A}$, where A is the class of all (t, s) pairs that represent appropriate term-sense associations.

Provided that the objects $x \in \mathcal{X}$ are represented in a suitable manner, most commonly as numerical *feature vectors* $\mathbf{x}$ in an m-dimensional Euclidean feature space $\mathbb{R}^m$, one of several learning algorithms can be employed to learn a classification. Support vector machines constitute a class of algorithms based on the idea of computing a decision hyperplane $\mathbf{w}^T\phi(\mathbf{x}) + b = 0$ that maximizes the margin between positive and negative training instances in the feature space (Vapnik 1998). Such maximum-margin hyperplanes tend to entail lower generalization errors than other separation surfaces, and the task of finding them leads to a quadratic optimization problem. Additional slack variables may be included to obtain a soft margin solution that is able to cope with training data that cannot be separated cleanly (Cortes and Vapnik 1995). The decision surface can be computed using Lagrange multipliers and decomposing techniques such as sequential minimal optimization (Platt 1999).

Using a simple dot product, we can then determine the distance $f(x) \in \mathbb{R}$ of a new instance x to this decision hyperplane in the feature space. A sigmoid function can be used to estimate posterior probabilities $P(y = A|x) = \frac{1}{1 + \exp(af(x) + b)}$ from these distances, where parameter fitting for a and b is performed using maximum likelihood estimation on the training data (Platt 2000; Lin et al. 2007). These posterior probabilities can be interpreted as confidence values $c_{(t,s)} = c_x = P(y = A|x)$ for a given instance $x = (t, s)$.

4.5 Generating the wordnet instance

We then apply one of the following rules for every (t, s) where t is an L_N-term from the translation dictionary and $s \in C(t)$ is a candidate sense as defined earlier:

(a) accept as a weighted connection with weight $c_{(t,s)}$ if and only if $c_{(t,s)} > 0$, or
(b) accept as an unweighted connection if and only if either $c_{(t,s)} \geq c_{\min}$, or $c_{(t,s)} \geq c'_{\min}$ and $\forall s' \neq s : c_{(t,s)} > c_{(t,s')}$ (for two pre-defined constants $c_{\min}$ and $c'_{\min} \leq c_{\min}$).

The first rule results in a weighted statistical wordnet for L_N, whereas the second one yields a more conventional unweighted wordnet.

Finally, new connections as well as of course new senses may be introduced manually to make the wordnet more complete. The introduction of new senses is particularly likely to be necessary for terms in L_N that had empty candidate sets.

Relational information for new synsets needs to be added manually. For the original synsets from the existing wordnet, we can immediately import a large number of links. Most importantly, hypernym links between synsets that have been found to have lexicalizations in L_{N} can quite safely be transferred to the new wordnet. It should however be noted that certain relations need to be re-interpreted as generic relatedness links between senses (e.g. the derivation relation), or are completely excluded from being imported (e.g. region domains). These issues are discussed in more detail in Sect. 5.4.

5 Evaluation and analysis of a machine-generated wordnet

While our approach is applicable to virtually any language, in the remainder of this article, we will focus on a German-language wordnet produced using our machine learning approach. Princeton WordNet 3.0, which covers around 155,000 English terms and around 118,000 senses, served as the existing template for the new wordnet. We further relied on the Ding German-English dictionary (Richter 2007), a large and fairly reliable digital translation dictionary with around 216,000 entries, but not much additional information apart from optional part of speech tags. A linear kernel SVM decision hyperplane was computed using LIBSVM (Chang and Lin 2001) and a training set consisting of 1,834 candidate associations (for 350 randomly selected German terms) that were manually classified as correct (22 %) or incorrect. The values $c_{\mathrm{min}} = 0.5$ and $c'_{\mathrm{min}} = 0.45$ were chosen as classification thresholds as described in Sect. 4 to generate the German wordnet. In order to obtain unbiased evaluation results, no form of manual revision was performed.

5.1 Accuracy and coverage

When evaluating the quality of this wordnet, we cannot rely on existing wordnets because these only provide positive examples but not negative ones, e.g. the fact that GermaNet (Kunze and Lemnitzer 2002) does not list the body of artists or thinkers sense of "*Schule*" (as in "*Frankfurter Schule*") does not imply that this sense association is incorrect. Instead, we considered a test set of 1,624 labelled sense associations obtained in the same way as the training set but completely independent from it, and thus not involved in any way in the wordnet building process. One can then evaluate to what degree the generated wordnet corresponds with the test set using standard evaluation measures. Given a test set, the precision is defined as $\frac{P_T}{P_T+P_F}$, and the recall is defined as $\frac{P_T}{P_T+N_F}$, where P_T, P_F, N_F are the number of true positives, false positives, and false negatives, respectively. Table 1 summarizes the results for our German wordnet, showing the precision and recall with respect to this test set.

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

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

Table 2 Alternative confidence thresholds

c_{min}	c'_{min}	Precision (%)	Recall (%)
0.90	0.80	94.21	34.03
0.90	0.60	91.50	41.79
0.70	0.60	87.50	52.24
0.60	0.50	83.90	59.10
0.50	0.45	81.11	65.37
0.40	0.35	73.64	72.54
0.35	0.25	70.53	80.00
0.30	0.25	67.32	82.39
0.20	0.15	55.93	90.15
0.10	0.05	40.41	94.93

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. Table 2 provides a sample of results obtained using alternative thresholds. For very high recall levels, an increased precision might not be realistic even with purely manual work, considering that Miháltz and Prószéky (2004) report an inter-annotator agreement of 84.73 % for such associations.

In addition to the recall scores in Table 1, which are based on the test set, Table 3 provides absolute numbers of terms covered by the German wordnet (using the classification thresholds $c_{\mathrm{min}} = 0.5$ and $c'_{\mathrm{min}} = 0.45$). While smaller than GermaNet 5.0, this automatically generated wordnet instance is already larger by an order of magnitude than many other manually compiled ones.

Table 4 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 that are polysemous with respect to the wordnet.

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*", "*Leydener Flasche*"), certain very common terms were found to be missing (e.g. "*Kofferraum*", "*Schloss*", "*Bank*"). This seems to arise from the fact that common terms tend to be more polysemous, thus making automatic associations difficult, though frequently such terms also have multiple translations, which tends to facilitate the mapping process. One solution would be manually adding associations for terms with

Table 3 Quantitative Assessment of Coverage of the German wordnet

	Sense associations	Terms	Lexicalized senses
Nouns	53,146	35,089	28,007
Verbs	13,875	5,908	6,304
Adjectives	21,799	13,772	9,949
Adverbs	4,243	2,992	2,593
Total	93,063	55,522	46,853

Table 4 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

high corpus frequency values, which due to Zipf's law would quickly improve the relative coverage of terms in ordinary texts. Another option is to rely on multilingual evidence (de Melo and Weikum 2009).

- Lexical Gaps and Incongruences: Another issue is the lack of senses for which there are no lexicalized translations in the English language, or which are not covered appropriately 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_N: Certain multi-word translations in L_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 or in corpus-derived lists.

Of course, the most general and reliable solution to ensure that the wordnet truly captures the typical senses of all terms and is free of incorrect sense associations is to perform a complete manual verification and revision process.

5.2 Comparison with alternative approaches

Our technique is further compared to four alternative approaches. We study the first sense heuristic, which involves simply accepting the first sense listed by WordNet for any English term. This heuristic is frequently cited as being more successful than many

Table 5 Comparison with existing methods

	Precision (%)	Recall (%)
First sense heuristic	40.36	67.46
Rigau & Agirre	48.97	63.58
Atserias et al.[a]	75.00	35.82
Benítez et al.	69.72	45.37
Our approach	81.11	65.37

[a] Excluding criteria based on additional background knowledge (see text)

other methods in word sense disambiguation tasks because the rank reflects the corpus frequency and importance of a sense. We also evaluate existing automatic approaches presented in Sect. 3. For Rigau and Agirre (1995), we considered the approach described in the second part of their paper, which was used to obtain a preliminary Spanish WordNet. From the study by Atserias et al. (1997), we consider the monosemy 1–4, variant, as well as the combined brother and polysemy 1/2 criteria. The CD criteria and the field criterion were not applied because their implementation in the original study is mainly based on additional lexical information for the Spanish language apart from the list of translations. The results, presented in Table 5, demonstrate that our learning-based approach outperforms the existing approaches both in terms of precision as well as in terms of recall. While two previous heuristics arrive at similarly high levels of recall, this occurs at the expense of very low precision scores. By adjusting the $c_{\min}, c'_{\min}$ confidence thresholds, our method can be made to produce recall scores well above 90 % at such levels of precision (cf. Table 2).

5.3 Relational coverage

By producing associations with senses of an existing source wordnet, we have the great advantage of immediately being able to import relations between the respective synsets. An excerpt of some of the relations we imported is given in Table 6.

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 7, which shows the results of a human evaluation for several different relation types. Incorrect relations are almost entirely due to incorrect term-sense associations.

5.4 Structural adequacy

As mentioned earlier, our machine learning approach is very parsimonious with respect to L_N-specific prerequisites, and hence scales well to new languages. Some might 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 or may not make the right sense distinctions, as in the case of the German "*Feierabend*". This point has already been discussed in Sect. 5.1. It

Table 6 An excerpt of some of the imported relations

Relation	Full links	Outgoing
Hyponymy	26,324	60,062
Hypernymy	26,324	33,725
Similarity	10,186	14,785
Has category	2,131	2,241
Category of	2,131	6,135
Has instance	641	5,936
Instance of	641	1,131
Part meronymy	2,471	6,029
Part holonymy	2,471	3,408
Member meronymy	400	734
Member holonymy	400	1,517
Substance meronymy	190	325
Substance holonymy	190	414
Antonymy (as sense opposition)	4,113	5,393
Derivation (as semantic similarity)	42,364	54,292

We distinguish full links between two senses both with L_N-lexicalizations, and outgoing links from senses with an L_N lexicalization

Table 7 Quality assessment for imported relations: For each relation type, 100 randomly selected links between two senses with L_N-lexicalizations were evaluated

Relation	Accuracy (%)
Hyponymy, hypernymy	84
Similarity	90
Category	91
Instance	93
Part meronymy, holonymy	83
Member meronymy, holonymy	89
Substance meronymy, holonymy	83
Antonymy (as sense opposition)	95
Derivation (as semantic similarity)	96

should also be clear that senses without any associated terms are to be considered unlexicalized nodes that do not directly represent the lexicon of the language.

Apart from these two considerations, it seems that general structural differences between languages rarely are an issue. When new wordnets are built independently from existing wordnets, many of the structural differences will not be due to actual conceptual differences between languages, but rather result from subjective decisions made by the individual human modellers (Pianta et al. 2002).

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.

6 Applications

6.1 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 sense associations. As people are more familiar with high-quality print dictionaries, they do not expect to encounter incorrect entries when consulting a WordNet-like resource.

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 6.1 with $n_h = 2$, $n_o = 2$, $n_g = 1$). Table 8 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 6.1 Thesaurus Generation

Input: a wordnet instance W (with function σ for retrieving senses and σ^{-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($W, n_h, n_o, n_g, R$)
    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 8 Sample entries from generated thesaurus (which contains entries for 55,522 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

6.2 Natural language processing

In this section, we will discuss some of the possible applications of automatically generated wordnets.

It turns out that the alignment with the English WordNet proves to be a major asset not only for cross-lingual but also 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 topical domain information, links to ontologies, and a range of other third-party resources described in more detail in Sect. 2.

For the task of word sense disambiguation, Patwardhan et al. (2003) presented an algorithm that maximizes the overlap of the English-language glosses (Patwardhan et al. 2003) with promising results, however we were unable to evaluate it more adequately due to the lack of an appropriate sense-tagged test corpus. One issue we noted was that the generated wordnet did not always cover all of the terms and senses to be disambiguated, which means that it is not a perfect sense inventory for word sense disambiguation tasks.

Apart from this, we believe that automatically generated wordnets are well-suited for virtually all other tasks that wordnets can been used for, including conventional information retrieval, multimedia retrieval, cross-lingual information retrieval (Chen et al. 2000), text classification, text summarization, coreference resolution (Harabagiu et al. 2001), machine translation, as well as semantic relatedness estimation and cross-lingual text classification, which we will now consider in more detail.

6.3 Case study: 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 (2005) reported that Lesk-style similarity measures based on the similarity of gloss descriptions (Lesk 1986) 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, one estimates their semantic relatedness using the maximum relatedness score between any of their two senses:

$$\mathrm{sim}(t_1, t_2) = \max_{s_1 \in \sigma(t_1)} \max_{s_2 \in \sigma(t_2)} \mathrm{sim}(s_1, s_2) \tag{7}$$

For the relatedness scores sim(s_1, s_2), we consider three different approaches, described in more detail in Sect. 4.3.2

1. $\mathrm{sim}_\mathrm{n}(s_1, s_2)$: graph neighbourhood proximity
2. $\mathrm{sim}_\mathrm{c}(s_1, s_2)$: cosine similarity of extended glosses
3. $\mathrm{sim}_\mathrm{m}(s_1, s_2)$: maximum (meta-method)

For evaluating the approach, we employed three German datasets (Gurevych 2005; Zesch and Gurevych 2006) 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 9, where we also list the current state-of-the-art scores obtained for GermaNet and Wikipedia as reported by Gurevych et al. (2007).

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.

Table 9 Evaluation of semantic relatedness measures, using Pearson's sample correlation coefficient

Dataset	GUR65		GUR350		ZG222	
	Pearson r	Coverage	Pearson r	Coverage	Pearson r	Coverage
Inter-Annot. Agreem.	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

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. (2007), one based on Wikipedia, the other on GermaNet

6.4 Case study: 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 occurrences 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 (2002).

In comparison with the standard monolingual case, 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 deliver poor results. 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 (Schmid 1994), 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 $\frac{w_{t,s}}{\sum_{s' \in \sigma(t)} w_{t,s'}}$ 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 Sect. 4.5. 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:

$$\log(n(t,d) + 1)\ \log\left(\frac{|D|}{|\{d \in D \mid n(t,d) \geq 1\}|}\right) \tag{8}$$

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 (Reuters 2000a, b). 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

Table 10 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

	Accuracy	Precision	Recall	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

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

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 (Joachims 1999) of support vector machine learning (Vapnik 1998), which is known to work very well for text classification.

The results in Table 10 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 wordnets can be built automatically if we are willing to accept a certain percentage of imprecise sense associations, and that these resources are nevertheless quite useful for various purposes. Our approach to constructing wordnets is based on statistical learning from a number of numeric scores and leads to a better coverage than the hard criteria proposed in previous studies, while simultaneously also allowing for a higher level of accuracy.

We have since conducted further experiments demonstrating that the method presented scales well to new languages (de Melo and Weikum 2009), as care was taken to require just a minimal amount of information specific to L_N. This enables us to produce a large-scale multilingual wordnet covering many different languages, available at http://www.mpii.de/yago-naga/uwn/.

Wordnets of this sort greatly facilitate interoperability, as they are aligned to the original Princeton WordNet, and thus also to other resources that are similarly aligned. First of all, of course, the machine-generated wordnets 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 revision 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 automatic techniques for extending the coverage of such statistically generated wordnets to senses not covered by the existing wordnets. We hope that our research has contributed to making lexical resources available for languages that previously had not been considered by the wordnet community.

References

Atserias, J., Climent, S., Farreres, X., Rigau, G., & Rodríguez, H. (1997). Combining multiple methods for the automatic construction of multilingual WordNets. In *Proceedings of the international conference on recent advances in NLP 1997* (pp. 143–149).

Baker, C., & Fellbaum, C. (2008). Can wordnet and framenet be made “interoperable”? In *Proceedings of the first international conference on global interoperability for language resources*.

Benitez, L., Cervell, S., Escudero, G., Lopez, M., Rigau, G., & Taulé, M. (1998). Methods and tools for building the Catalan WordNet. In: *Proceedings of the ELRA workshop on language res. for Europ. Minority Lang., 1st international conference on language resources and evaluation*.

Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising the WordNet domains hierarchy. In *COLING 2004 multiling. Ling. Resources, Geneva, Switzerland* (pp. 94–101).

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—the story so far. *International Journal on Semantic Web and Information Systems, 5*(3), 1–22.

Buscaldi, D., & Rosso, P. (2008). Geo-wordnet: Automatic georeferencing of wordnet. In (ELRA) ELRA (Ed.), *Proceedings of the 6th international language resources and evaluation (LREC’08)*, Marrakech, Morocco.

Chang, C. C., & Lin, C. J. (2001) LIBSVM: A library for support vector machines. URL http://www.csie.ntu.edu.tw/cjlin/libsvm.

Chen, H. H., Lin, C. C., & Lin, W. C. (2000). Construction of a Chinese-English WordNet and its application to CLIR. In *Proceedings of the fifth international workshop on information retrieval with Asian languages, IRAL ’00* (pp. 189–196). New York, NY, USA: ACM Press.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Cycorp Inc. (2008). Opencyc. http://www.opencyc.org/.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Li, F. F. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR 2009)*.

de Melo, G., & Siersdorfer, S. (2007). Multilingual text classification using ontologies. In G. Amati (Ed.), *Proceedings of the 29th European conference on information retrieval (ECIR 2007)*. Springer, Rome, Italy, *Lecture Notes in Computer Science*, Vol. 4425.

de Melo, G., & Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on information and knowledge management (CIKM 2009)* (pp. 513–522). New York, NY, USA: ACM.

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database (language, speech, and communication*. Cambridge: The MIT Press.

Francopoulo, G., Declerck, T., & Sornlertlamvanich, V., de la Clergerie, E., & Monachini, M. (2008). Data category registry: Morpho-syntactic and syntactic profiles. In *Proceedings of the workshop on use and usage of language resource-related standards at the LREC 2008*.

Gangemi, A., Navigli, R., & Velardi, P. (2003). The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet. In *On the move to meaningful internet systems 2003: CoopIS, DOA, and ODBASE* (pp. 820–838).

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the second international joint conference on natural language processing, IJCNLP, Jeju Island, Republic of Korea*.

Gurevych, I., Müller, C., & Zesch, T. (2007). What to be?— electronic career guidance based on semantic relatedness. In *Proceedings of the 45th annual meeting of the association for computational linguistics, Association for Computational Linguistics, Prague, Czech Republic* (pp. 1032–1039).

Harabagiu, S. M., Bunescu, R. C., & Maiorano, S. J. (2001). Text and knowledge mining for coreference resolution. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001, Association for Computational Linguistics, Morristown, NJ, USA* (pp. 1–8).

Joachims, T. (1999). Making large-scale support vector machine learning practical. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods: Support vector machines*. Cambridge, MA, USA: MIT Press.

Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. In *AAAI* (pp. 691–696).

Knight, K. (1993). Building a large ontology for machine translation. In *Proceedings of the workshop human language technology* (pp. 185–190).

Kunze, C., & Lemnitzer, L. (2002). GermaNet—representation, visualization, application. In *Proceedings of the LREC 2002* (pp. 1485–1491).

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, SIGDOC '86* (pp. 24–26). New York, NY, USA: ACM Press.

Lin, H. T., Lin, C. J., & Weng, R. C. (2007). A note on platt's probabilistic outputs for support vector machines. *Machine Learning, 68*(3), 267–276.

Lyons, J. (1977). *Semantics, Vol. 1*. Cambridge: Cambridge University Press.

Miháltz, M., & Prószéky, G. (2004). Results and evaluation of Hungarian Nominal WordNet v1.0. In *Proceedings of the second global WordNet conference*. Brno, Czech Republic: Masaryk University.

Niles, I., & Pease, A. (2003). Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In *Proceedings of the 2003 international conference information and knowledge engineering, Las Vegas, NV, USA*.

Okumura, A., & Hovy, E. (1994). Building Japanese-English dictionary based on ontology for machine translation. In *Proceedings of the workshop on human language technology* (pp. 141–146).

Ordan, N., & Wintner, S. (2007). Hebrew WordNet: A test case of aligning lexical databases across languages. *International Journal of Translation, 19*(1), 39–58.

Patwardhan, S., Banerjee, S., & Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. In *Proceedings 4th international conference on computational linguistics and intelligent text processing (CICLing), Mexico City, Mexico*.

Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st international global WordNet conference, Mysore, India* (pp. 293–302).

Platt, J. C. (1999). *Fast training of support vector machines using sequential minimal optimization* (pp. 185–208). Cambridge, MA, USA: MIT Press.

Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). Cambridge, MA, USA: MIT Press.

Reuters. (2000a). Reuters Corpus, Vol. 1: English language, 1996-08-20 to 1997-08-19. URL http://trec.nist.gov/data/reuters/reuters.html.

Reuters. (2000b). Reuters Corpus, Vol. 2: Multilingual, 1996-08-20 to 1997-08-19. http://trec.nist.gov/data/reuters/reuters.html.

Richter, F. (2007). Ding version 1.5. http://www-user.tu-chemnitz.de/~fri/ding/.

Rigau, G., & Agirre, E. (1995). Disambiguating bilingual nominal entries against WordNet. In *Proceedings of the Workshop 'The Computational Lexicon' at European summer school logic, language & information*.

Sathapornrungkij, P., & Pluempitiwiriyawej, C. (2005). Construction of Thai WordNet lexical database from machine readable dictionaries. In *Proceedings of the 10th machine translation summit, Phuket, Thailand.*

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International conference on new methods in language processing, Manchester, UK.*

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A Core of semantic knowledge. In *16th International World Wide Web Conference (WWW 2007)*. New York: ACM Press.

Tufiş, D., Ion, R., & Ide, N. (2004). Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *COLING '04: Proceedings of the 20th international conference on computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA* (p. 1312).

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley-Interscience.

Vossen, P. (Ed.) (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Berlin: Springer.

Zesch, T., & Gurevych, I. (2006). Automatically creating datasets for measures of semantic relatedness. In *COLING/ACL 2006 workshop on linguistic distances, Sydney, Australia* (pp. 16–24).