Extraction, Evaluation and Integration of Lexical-Semantic Relations for the Automated Construction of a Lexical Ontology

Tonio Wandmacher, Ekaterina Ovchinnikova, Ulf Krumnack, Henrik Dittmann

Artificial Intelligence group
Institut für Kognitionswissenschaft
Universität Osnabrück, Germany
Email: {firstname.lastname}@uni-osnabrueck.de

Abstract

Several approaches for extracting semantic relations from various types of resources have been proposed during the last years. While already of great value when used separately, combining these techniques promises to lead to even broader and more reliable results. However, divergent information may occur when assembling such data. We present LexO, a framework for integrating semantic relations from different sources into an ontological structure. We provide different methods for assigning confidence values to the input data as well as mechanisms to detect and resolve inconsistencies. The present paper focuses on lexical-semantic relations, but the approach presented is extensible to include new kinds of data sources as well as further types of relations.

1 Introduction

The construction of ontologies is considered essential not only in the development of the semantic web but also for a growing number of natural language processing (NLP) tasks such as word sense disambiguation, automatic semantic annotation of documents, question answering, machine translation and anaphora resolution.

Whereas most ontologies are constructed for a given domain and contain relations between concepts, a *lexical ontology* is intended to provide structured information on words of a given language and their semantic relatedness; meaning is encoded by relating a given lexical item to others. Also, the main goal of a lexical ontology is not to store general encyclopedic or ontological knowledge, but to serve as common database, assembling lexical and semantic information.

In the past years a number of projects have been presented that try to achieve this goal, of which the most prominent one is the Princeton WordNet (Fellbaum 1998). It represents domain independent, lexical-semantic knowledge in a network-like structure which makes taxonomic relationships explicit. However, it cannot be considered as an ontology in the formal sense, since the relations are based on linguistic evidence rather than on formal ontological principles, and it does not guarantee any kind of consistency (cf. (Oltramari et al. 2002) for examples of ontological inconsistencies in WordNet).

The main problem, however, remains data coverage. Even though WordNet and its cousins are considered as broad coverage resources, many NLP applications run into problems of data sparsity when relying on such resources only, which are all developed manually at great cost. A possible solution to the sparsity problem present automatic extraction procedures. In the past years a lot of automated approaches have been presented to extract ontological knowledge from text or even structured data (for an overview cf. Maedche 2002 or Cimiano 2006). The main problem of these approaches however is their reliability; as every unsupervised procedure, they also extract noise. A way to overcome the problems of low coverage and low data quality is the cumulation of evidence. When many available resources and extraction procedures are exploited at the same time, reliable relations can be distinguished from noise, if practical measures to estimate the confidence of each relation are provided.

To our knowledge however, no bigger attempt has been made to realize this idea. The approach described by Cimiano et al. (2005) has gone in this direction by integrating taxonomic relations from different ontology learning paradigms. Another interesting work (Snow et al. 2006) presents an algorithm to induce a domain-independent taxonomy from heterogenous resources by defining several constraints on the resulting structure. However these approaches only consider *is-a* relations, and they have only been applied on a small scale.

The LexO project that we present here, aims at integrating any kind of lexical-semantic relation from automated extraction procedures and already existing, freely accessible lexical resources. Information from various origins is cumulated and integrated in a way which makes it possible to identify reliable relations. These relations will form a set of *hypotheses* from which an ontology is constructed. Our approach is highly automated. We present an elaborate measure to estimate the confidence for each incoming relation hypothesis. Our confidence measure takes into account the *a priori* confidence of the respective resource, semantic similarity between the connected terms and structural evidence from the already existing data. The ontology construction itself is automated as well, we define structural consistency conditions which have to be assured by the ontology to be constructed from the assembled relation hypotheses.

Although our work is focused on the creation of a lexical ontology for the German language, the overall approach is in principle language neutral: Methods to extract semantic relations have of course to be designed for an individual language, but they can easily be adapted to other languages. There might also exist other lexical-semantic resources to exploit; our framework takes advantage of any lexical resource and extraction method, as long as it can provide binary relations. Moreover, the types of relation are not fixed either; every relation can be modeled as long as

Copyright ©2007, Australian Computer Society, Inc. This paper appeared at the 3rd Australasian Ontology Workshop (AOW2007), Gold Coast, Queensland, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 85, Thomas Meyer and Abhaya Nayak, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

a resource is able to provide it.

At present, LexO comprises 975,570 relation entries (synonymy, hyponymy, meronymy and antonymy) over 121,593 unique words (types). So far, we make use of the following resources: *Wiktionary*, *OpenThesaurus* (Naber 2005), *Projekt Deutscher Wortschatz*[1], an (unsupervised) translation of WordNet and an automatic extraction method, looking for lexico-syntactic patterns on the web (similar to Cimiano & Staab 2004).

Since the LexO project is in an early stage of development, we cannot present an overall evaluation of our methods and the hereby constructed ontology yet. The aim of this paper is to present measures to evaluate the confidence of automatically extracted lexical-semantic relations and to describe a way to integrate these relations in a consistent manner.

The paper is structured as follows: In section 2 we give an overview on methods and resources providing lexical-semantic relations, we then (section 3) describe measures to estimate the confidence of these relations, in section 4 we deal with the problem of word senses and formulate consistency conditions for the resulting ontological structure, and in section 5 we present the overall architecture of the LexO system and describe possible evaluation scenarios. In the final section we then discuss open issues and describe the following steps of our work.

2 Obtaining semantic relations

Semantic relations between some items are relations between meanings of this items; lexical-semantic relations are thus relations between meanings of words (cf. Cruse 1986). The term *lexical ontology* (LO) is rather underspecified in the existing literature. Usually it means that words of a particular language (rather then abstract concepts) are formally defined and connected with each other by lexical-semantic relations such as *synonymy*, *hyponymy* or *meronymy*. WordNet is considered to be the most typical example of LO. In the LexO framework, a lexical ontology is a set of relations over a domain of words or word senses (unlike WordNet, where relations can hold between synsets). Every relation is a set of pairs of objects from the domain.

While LexO aims at collecting various kinds of relations, this paper focuses on lexical-semantic relations, i.e. relations that are founded on the meaning of words rather than on their form. This section describes different techniques to obtain such relations from various resources.

2.1 Existing approaches in ontology learning

In the past few years a variety of approaches has been presented that aim at extracting conceptual knowledge from unstructured and semi-structured data. These approaches receive a growing importance in the ontology building process, since for many semantic web as well as NLP applications the amount of available knowledge is crucial. Since these methods are unsupervised, their output is usually rather noisy.

So far, most of the approaches are light-weight from a logical point of view; they return logically simple constructions such as concepts, instances, taxonomic relations and other general relations (e.g. *part-of* or *author-of*). Current methods basically make use of three strategies (or combinations of these):

1. *Distributional information:* The co-occurrence of terms within a given context or document is an important hint for their conceptual relatedness. Moreover, two terms will be similar in meaning if they tend to occur with the same neighbors (2nd order cooccurrence). Different distributional methods (e.g. collocation analysis or *Latent Semantic Analysis*, Deerwester et al. 1990) give a distance measure between two terms that can be used to represent semantic relatedness. Even though this cannot help labeling the type of relation, it gives a reliable clue that can be further used. Clustering techniques for example use this information to form sets of related terms. In hierarchical clustering procedures, these sets of terms are arranged in a hierarchical fashion. The hereby generated cluster hierarchy can be the base for a taxonomical structure, i.e. a hierarchy of concepts. Approaches that use this kind of strategy are for example described by Caraballo (1999) or Cimiano & Staab (2005).
2. *Lexico-syntactic patterns:* The second strategy basically relies on lexico-syntactic patterns, the so-called *Hearst* patterns (Hearst 1992). Here, a text corpus is scanned for characteristic recurring word combinations, typically containing a semantic relation between two terms (e.g. $[w_2$, `such as` $w_1] \rightarrow hyponym(w_1, w_2)$). These approaches however usually suffer from data sparsity, since many word combinations cannot be found in even large corpora. To cope with this fact, efforts been made to harvest these patterns on the web (cf. Brin 1998, Etzioni et al. 2004 or Cimiano & Staab 2004).
3. *Syntactic and morphosyntactic information:* Finally linguistic structures like verb frames and modifier constructions can help extracting conceptual relations. For example, it is easy to infer a hyponymy relation between *car ferry* and *ferry*, since *car* is here a modifier of *ferry* (cf. Buitelaar et al. 2004). Moreover, from the analysis of dependency paths in syntactic derivations, reliable relations can be learned (Katrenko & Adriaans 2006), other methods make use of predicate-argument relations (e.g. Faure & Nédellec 1998). For the extraction of nontaxonomic relations the analysis of selectional preferences of verbs can be very helpful (Wagner 2000).

Techniques based on these strategies can be found in many ontology learning systems, such as *Snowball* (Agichtein & Gravano 2000), *OntoLearn* (Navigli & Velardi 2004), *OntoLT* (Buitelaar, Olejnik & Sintek 2004), and *Text2Onto* (Cimiano & Völker 2005). Most of these systems are concerned with the extraction of the relevant terminology (from which they deduce the respective classes), with the derivation of subsumption relations and with some basic nontaxonomic relations.

2.2 Automatic translation of *WordNet*

A rich source of relational lexical information are wordnets, especially the English WordNet. WordNet represents knowledge in form of a lexical network. Its organizing units are sets of synonyms (so-called *synsets*), representing word meanings. Two kinds of relations can be distinguished: a) relations connecting individual lexical items and b) relations connecting synsets and thus providing a statement indirectly via the synonymy relation. Both kinds of relations can be used as input for LexO.

Although nowadays wordnets exist for many languages,[2] their benefit often is restricted due to lim-

[1] http://wortschatz.uni-leipzig.de/

[2] A current list is maintained by the Global WordNet Association at `http://www.globalwordnet.org/gwa/wordnet_table.htm`

ited size or license issues, like the German GermaNet (Hamp & Feldweg 1997), which is protected. Therefore in our context, a translation of the English WordNet can be a promising alternative. There have been a number of approaches that use bilingual dictionaries to apply automatic and semi-automatic methods to translate WordNet into different languages (e.g. Spanish (Knight & Luk 1994), Japanese (Okumura & Hovy 1994) or Arabic (Khan & Hovy 1997)). Most problems in such approaches are caused by polysemy, mismatches between the bilingual dictionary and WordNet, as well as mismatches in the lexicalization between the languages.

Various techniques have been proposed to deal with ambiguities that arise when mapping dictionary entries to WordNet synsets (cf. Atserias et al. 1997). They are based on additional information from WordNet and the dictionary such as part-of-speech, alternative translations, domain markers, syntactic and semantic annotation or frequency information. Consider for example the synset

$$\{plant,\ flora,\ plant\ life\}$$

and a dictionary entry of the form

$$plant \rightarrow Pflanze\ [bot.];\ Werk$$

There are two different translations for *plant*,[3] but as *plant* is polysemous in WordNet, it is not clear, which translation should be mapped to the synset. Here a human can use the domain marker *[bot.]* to disambiguate the translation. To make this strategy available for automatic translation methods, WordNet has to be annotated with the domain markers of the dictionary, a feasible task, as there are usually only few domain markers in use.

An alternative strategy to translate synsets with more than one element is to collect the translations for every word in the synset and consider their intersection. In the above example this means to look at the following entries:

$$flora \rightarrow Flora,\ Pflanzenwelt\ [biol.]$$
$$plant\ life \rightarrow Pflanzenwelt$$

Here *Pflanzenwelt* seems to be a promising translation for the synset (this assumption is further strengthened by the fact that *plant life* is monosemous in WordNet). However, in many cases the intersection is empty, but there are translations that are semantically similar. Given a measure for semantic similarity of words in the target language, this can be used in cases when a common translation is missing.

In most cases such disambiguation techniques do not lead to a definitive selection but rather rank the alternatives. In the context of our work, such a ranking can be used to assign a confidence score to induced relation hypotheses.

There is some agreement that an automatic translation will not result in a ready-to-use WordNet for the target language. However, for our approach, relations stemming from such a translation process, annotated with confidence values, are valuable input material. Once an initial lexical ontology is constructed for the target language, it can be used to foster the disambiguation process, providing in turn more confident hypotheses.

2.3 Obtaining relations from electronic dictionaries and thesauri

In recent years, many lexical resources have been made electronically available. A lot of these provide free access over the internet and often have liberal licenses governing their use and re-distribution. We present three examples for German.

The *Wiktionary* project is an offshoot of *Wikipedia*[4], the well-known open encyclopedia. Online since 2002, the site provides dictionaries for a large number of languages. Each of these may contain entries from any language, which are explained in the language of the respective dictionary. Like its sister project, *Wiktionary* is a collaborative effort where basically everyone can participate in its construction. Often such a dictionary's base is assembled by automatic extraction from other publicly available sources, however. The German *Wiktionary* has been online since 2004 and currently has 55,000 entries for all languages, of which more than 40% are for German words.

A *Wiktionary* entry for a given word may comprise all kinds of lexical information, such as phonetics, morphological properties, etymology, word senses and semantic relations (e.g., synonyms, antonyms and hypo-/hyperonyms). At present, we extract all lexical-semantic relations between German words that can be identified through the page structure and markup, taking note of word senses whenever they are present in the resource.

The project *OpenThesaurus* (Naber 2005) has been online since 2003. A freely accessible and modifiable resource for the German language, *OpenThesaurus* is primarily structured through groups of synonyms. The project aims at organizing these groups in a hierarchical WordNet-like (Fellbaum 1998) manner, starting from a small range of top-level concepts. In doing this, hypo-/hyperonym relationships are added between the synonym groups. However, to date only a fraction of groups have been attached to the hierarchy. *OpenThesaurus* provides its data in a variety of formats, such as a plain database dump or a plug-in to *OpenOffice*.

The *Projekt Deutscher Wortschatz*[5] at the Universität Leipzig is a monolingual German dictionary, comprising more than 9 million full (i.e., inflected) forms and multi word units. The dictionary is largely based on automatic extraction methods for corpora in conjunction with reviewing and editing by human experts and has more restrictive terms of use than the previous examples.

For a given word, information is provided on grammatical status, frequency, topical domain(s) and semantic relations. Example phrases and automatically calculated co-occurrences and collocations are provided as well. This data is available through either a web interface or a number of web services for automated retrieval.

The example in table 1 shows the relations for the noun *Stern* ('star'), as extracted from the resources mentioned above.

3 Calculating confidence

Estimating the reliability of a given relation is a non-trivial problem for an automated approach, but it is crucial to have such a measure in order to build up an ontology of high quality. In the following we show how we calculate our confidence scores, which are comprised of a local confidence value for a given relation as provided by its resource, the overall reliability of its resource, structural criteria and an automatically calculated similarity score. For this purpose we make use of *Latent Semantic Analysis* (LSA), a vector-based method which has been shown to give reliable estimates on semantic similarity.

[3] *Pflanze*: 'botanical plant'; *Werk*: 'factory'/'work'

[4] http://www.wikipedia.org

[5] http://wortschatz.uni-leipzig.de

OpenThesaurus	Wiktionary	Wortschatz Projekt
synonym($Stern_1$, *Asterisk*)	*hyponym*($Stern_a$, *Himmelskörper*)	*synonym*(*Stern*, *Filmstar*)
synonym($Stern_1$, *Asteriskus*)	*synonym*($Stern_a$, *Gestirn*)	*synonym*(*Stern*, *Gestirn*)
synonym($Stern_1$, *Sternchen*)	*synonym*($Stern_a$, *Fixstern*)	*synonym*(*Stern*, *Star*)
hyponym($Stern_2$, *Gestirn*)	*hyponym*($Stern_b$, *Symbol*)	*synonym*(*Stern*, *Planet*)
hyponym($Stern_2$, *Himmelskörper*)	*synonym*($Stern_b$, *Asterisk*)	*hyponym*(*Stern*, *Gestirn*)
synonym($Stern_2$, *Fixstern*)	*synonym*($Stern_b$, *Sternchen*)	*hyponym*(*Stern*, *Himmelskörper*)
synonym($Stern_3$, *Star*)	*hyponym*($Stern_c$, *Mensch*)	*hyponym*(*Stern*, *Schmuck*)
...	*hyponym*($Stern_c$, *Kosewort*)	...
...	...	...

Table 1: Relations for *Stern* ('star') from *Wiktionary*, *OpenThesaurus* and *Wortschatz*.

3.1 LSA-based semantic similarity

Since the early 1990s, Latent Semantic Analysis (LSA) has become a well-known technique in NLP. When it was first presented by Deerwester et al. (1990), it aimed mainly at improving the vector space model in information retrieval, but in the meantime it has become a helpful tool in NLP as well as in cognitive science (cf. Landauer & Dumais 1997). LSA has been shown to give reliable estimates for the semantic similarity between two terms, and it has also been used to enhance automatic hyponymy extraction techniques (Cederberg & Widdows 2003). If two terms receive a high LSA similarity value, they will be somehow semantically related, however LSA cannot determine the kind of relation (Wandmacher 2005).

The LSA model is based on the vector space model from information retrieval (IR) (Salton & McGill 1983). Here, a given corpus of text is first transformed into a term×context matrix A, displaying the occurrences of each word in each context. Usually, this matrix is then weighted by one of the standard weighting methods used in information retrieval (c.f. Salton & McGill 1983). The decisive step in the LSA process is then a *singular value decomposition* (SVD) of the weighted matrix. Thereby the original matrix A is decomposed as follows:

$$SVD(A) = U\Sigma V^T \tag{1}$$

The matrices U and V consist of the eigenvectors of the columns and rows of A. Σ is a diagonal matrix, containing in descending order the singular values of A. By only keeping the k strongest (k usually being 100 to 300) eigenvectors of either U or V, a so-called semantic space can be constructed for the terms or the contexts, respectively. Each term or each context then corresponds to a vector of k dimensions, whose distance to others can be compared by a standard vector distance measure. In most LSA approaches the *cosine* measure is used.

We use a slightly different setting, close to the one described by Schütze (1998) and Cederberg & Widdows (2003), where the original matrix is not based on occurrences of terms in documents, but on other cooccuring terms (term×term-matrix). We thus count the frequency with which a given term occurs with others in a predefined context window ($\pm 10 - 100$ words). After applying *singular value decomposition*, each word is represented as a vector of k dimensions, and for every word pair w_i, w_j of our vocabulary we can calculate a similarity value $Sim(w_i, w_j)$, based on the *cosine* between their respective vectors.

3.2 Local resource confidence (LRC)

When combining relations from different sources, not all of them will be equally reliable. Depending on the type of resource in question, relations can be already equipped with a confidence value. For example, an extraction technique matching lexico-syntactic patterns on the web counts the number of matches for two words w_i and w_j and a given pattern π ($[w_i\ \pi\ w_j]$). When this value is normalized by the maximum frequency of $[w_i\ \pi]$, each extracted relation triple t_k in resource r can be assigned a local resource confidence value $LRC(t_{kr})$ between 0 and 1. The following list defines the *local* confidence ratings that we use for the resources incorporated so far:

- *Wiktionary*: relative frequency of relation (frequency of a relation / number of relations)
- *OpenThesaurus*: relative frequency of relation
- *Wortschatz*: relative frequency of relation
- *Transl. WordNet*: mean translation confidence. Given a translation t_1 for a WN synset s_1 with a reliability $r_1 \in [0, 1]$ and a translation t_2 for synset s_2 with a reliability r_2, and given $R(s_1, s_2)$ in WordNet we set the *local resource confidence* of $R(t_1, t_2)$ to the mean of r_1 and r_2.
- *Hearst patterns*: maximum likelihood. Given two terms w_1 and w_2, matched by a pattern π ($w_1\pi w_2$), we divide the matching frequency of $w_1\pi w_2$ with the frequency of $w_1\pi$

Even though ranging between 0 and 1, we acknowledge that the mathematical properties as well as the semantics of these measures are difficult to compare. However, we prefer to exploit the confidence ratings provided by the resources themselves than to assume uniform confidence for every incoming relation.

3.3 Global resource confidence (GRC)

A hand-coded resource like *Wiktionary* is surely more trustworthy than automated extraction techniques, which yield usually rather noisy results. A relation coming from *Wiktionary* should therefore receive a higher overall confidence than one coming from a pattern-based approach. Estimating the overall confidence of a resource can be done by determining the average LSA similarity for all n word pairs w_i, w_j figuring in the relation triples t_k of the resource r. A high GRC value (formula 2) indicates that the terms connected via relations in that resource fall into one semantic field in real life texts. Table 2 shows GRC values for the resources we have integrated so far in LexO. A reference LSA space was calculated on a 101 million word corpus consisting of German *Wikipedia* and newspaper articles from a German daily (*Die Tageszeitung*, 1996 – 1999) and then reduced to 150 dimensions. For the calculation we used the *Infomap* toolkit, v0.8.6[6], the co-occurrence window was set to ± 100 words.

$$GRC_r = \frac{1}{n}\sum_{k=0}^{n} Sim_{LSA}(w_i, w_j) \tag{2}$$

[6] http://infomap-nlp.sourceforge.net/

Source	raw	norm.	human	Dev.
Wiktionary	0.163	0.74	70%	+4%
OpenThes.	0.147	0.68	74%	-6%
Hearst-p.	0.109	0.51	57%	-6%
transl. WN	0.087	0.41	40%	+1%
Wortschatz	0.138	0.62	40%	+22%

Table 2: LSA based confidence values and human judgements for different resources.

To evaluate the accuracy of the calculated *GRC* values, we drew for each resource a random test sample of 100 triples. These triples were manually evaluated by 3 human annotators. We asked the annotators simply to label if the given relation holds or not (e.g. "Is X a hyponym for Y?"). The percentages of correct relations, as judged by the annotators, are also given in table 2.

As can be seen immediately, these results correlate strongly with the *GRC* values, with one exception: The *GRC* value of the *Wortschatz* data is obviously overestimated by our automatic measure. This is due to an apparent weakness of LSA, which is not able to distinguish between the relation types. Further manual inspection showed that the *Wortschatz* data contain mostly relations which would more appropriately be labeled as "near"-synonyms or loose associations, not as true synonyms. The goal of our project is to rely as little as possible on manual human inspection, but so far, our *GRC* measure has no means to detect relation mislabeling. For this reason we use meanwhile for the *Wortschatz* data a corrected *GRC* value (0.40), and we will try to develop more sophisticated measures in order to better estimate the reliability of a resource.

3.4 Confidence from structural information

For the estimation of confidence for a given relation we can not only exploit information inherent to the relation and its resource, but also on evidence from the already assembled data. One would probably assume that a *synonym* relation (x, y) is more reliable, if we have already the inverse relation (y, x) in the data base. Likewise, if we find for a given *hyponym* relation (x, y) its inverse *hypernym* pair (y, x) (this counts also for *mero-* and *holonyms*), we want to give it a higher confidence rating. Finally, due to the (normally assumed) transitivity of hyponymy, if we find for a *hyponym* pair (x, y) also the *hyponym* pairs (y, z) and (x, z), we can assume (x, y) to be more reliable.

To make use of this kind of information, we define a range of indicator functions I_{1-4} returning 1, if one of the following conditions holds for a given triple $R(x, y)$, and 0 else.

1. **Synonym symmetry:**
   $I_1 = syn(x, y) \wedge syn(y, x)$
2. **Hypo-/hypernym correspondence:**
   $I_2 = hypo(x, y) \wedge hyper(y, x)$
3. **Mero-/holonym correspondence:**
   $I_3 = mero(x, y) \wedge holo(y, x)$
4. **Hypernym commonness:**
   $I_4 = hypo(x, y) \wedge hypo(x, z) \wedge hypo(y, z)$

The list of indicator functions is not meant to be exhaustive, there might be many more of such conditions playing a role in confidence estimation.

3.5 Individual semantic similarity

As long as we regard semantic relations, we can assume that the terms w_{k1} and w_{k2} of a triple t_k have a high semantic similarity as calculated by a method like *LSA*. This gives us another confidence measure for a given triple t_k:

$$Sim(t_k) = \nu \cdot (cos_{LSA}(w_{k1}, w_{k2})) \quad (3)$$

The factor ν normalizes the result, so that it also ranges between 0 and 1.

3.6 Integrated confidence

When we integrate the resources, we combine all single confidence values by linear interpolation. The *LRC* values ($LRC(t_{kr})$) of all resources for a relation are accumulated, according to the overall confidence GRC_r of the respective resource r. We then add the structural confidence and the semantic similarity score *Sim*.

$$\begin{aligned} IC(t_k) \quad = \quad & \lambda_1 \cdot \left(\nu \sum_{r=0}^{n} GRC_r \cdot LRC(t_{kr}) \right) \\ & + \lambda_2 \cdot I_1(t_k) \\ & + \lambda_i \cdot I_j(t_k) \ldots \\ & + \lambda_m \cdot Sim(t_k) \end{aligned} \quad (4)$$

After integration, every relation triple t_k has an integrated confidence value *IC*, calculated from the single confidence values of the resources, where t_k appeared, weighted by their respective *GRC* value, the structural confidence functions $I_i(t_k)$ and the similarity function $Sim(t_k)$. $\lambda_{1\ldots n}$ are the coefficients controlling the importance of each component and sum up to 1. They can be optimized by an *EM*-style algorithm (cf. Dempster et al. 1977). ν is a normalizing factor, assuring that the accumulated confidence scores remain between 0 and 1 and n the number of resources integrated so far.

4 Syntactic integration

After a new set of relation hypotheses has been collected from external sources, these data have to be added to the already cumulated lexico-semantic resource (which is empty in the first iteration). In this step we have to solve two main problems in order to create an integrated and consistent data set: unification of word senses and resolution of possible inconsistencies.

4.1 Dealing with word senses

One of the major problems in combining lexical data from different resources lies in the discrimination of word senses (WS). If the only identifier of a term is its lexical form, it is impossible to automatically distinguish polysemous words. This is not only impractical for many applications, it also leads to weird constructions in the resulting ontology. Suppose a data set contains the following triples:

hyponym(*Tree, Plant*)
hyponym(*Tree, Structure*)
hyponym(*Oak, Tree*)

Due to the transitivity of the relation *hyponym* an automatic reasoner would infer here that an oak is both a plant and a structure. Obviously, the identifier *Tree* needs to be split (e.g. $\texttt{Tree}_1$ for the *plant* sense, and $\texttt{Tree}_2$ for *structure*).

Fortunately, some of the resources that we are using (e.g. *Wiktionary* and *OpenThesaurus*) do distinguish WS, but most other data sets (esp. from automatic extraction methods) do not support WS distinction.

This problem of data integration is close to the problem of mapping from a lexical resource to an ontology (or to another lexical resource). This issue is discussed in the literature (cf. Niles & Pease 2003), however, no general mapping strategy is available. In LexO, we use corpora-based methods and contexts of terms in data sets for WS disambiguation. We define a *context* for a term t in a resource r as a set of all terms[7] that co-occur with t in triples from r (or co-occur with terms that co-occur with t). A similar method was used for example by Buitelaar & Sacaleanu (2001). The transitivity of hyponymy and meronymy is used to extend a context of a term t with all "ancestors" of t. The word senses of terms in the context sets are ignored, because every context set is supposed to define proper WS of its members.

Given a set of triples S_1 where the word senses are distinguished, another set of triples S_2 has to be integrated with S_1. Let us first consider the case when S_2 distinguishes between word senses. We illustrate this case by examples from *Wiktionary* and *OpenThesaurus*, presented in table 1. The term *Stern* ('star') is polysemous in both resources. The relations of this term are used to build its context. The context of $Stern_1$ in *OpenThesaurus* is *Asterisk, Asteriskus, Sternchen* and the context of $Stern_b$ in *Wiktionary* is *Asterisk, Sternchen, Symbol*. Since these contexts overlap (*Asterisk, Sternchen*), they are supposed to define the same WS. Thus, $Stern_1$ and $Stern_b$ are unified to $Stern_1$ in the resulting integrated data set. Resources may contain not enough information for word sense unifying (e.g. for $Stern_3$ and $Stern_c$ in our example). In this case it is necessary to refer to external information sources (cf. for example Dorow & Widdows 2003), or a method like LSA (cf. 3) providing a similarity measure for the contexts of $Stern_3$ and $Stern_c$.

If a resource to be added does not distinguish between word senses, then every term from this set has to be considered as potentially polysemous. Let us consider the triples extracted from our *Wortschatz* data. We cannot use the information about the combined context of *Stern* anymore and have to treat every triple separately. For example, if a triple *synonym*(*Stern, Filmstar*) ('star', 'movie star') is to be added, the context of *Stern* in this case will be limited to *Filmstar*. Again, an LSA-based method can be used to measure the similarity between the term *Filmstar* and the contexts of *Stern* in the integrated data set (*Asterisk, Symbol, ...* ('asterisk' , 'symbol'), *Himmelskörper, Gestirn, ...* ('heavenly body' sense) and *Mensch, Star* ('person' sense)).

4.2 Formulating Consistency Conditions

An important benefit of using a formalized ontological database in applications is the possibility to reason over the content of the ontology. For example, the inference of a subsumption hierarchy may help in formulating selectional restrictions, disambiguation tasks etc. But if the ontology contains mistakes and inconsistencies, reasoning may appear to be misleading and therefore pointless. There is a lot of literature on logical inconsistencies in ontological knowledge bases (cf. Kalyanpur 2006). However, as far as we know, no consistency constraints have been formulated yet for lexical resources (such as WordNet).

As we do not make use of complex logical statements (such as number restrictions, role inclusion, etc.) our resulting ontology is simple from a logical point of view.[8] Still, it should obey certain structural criteria: For example, we do not want to allow that two or more semantic relations hold between a term pair (e.g. *synonym*(w_1, w_2) and *hyponym*(w_1, w_2)). Another structure that should be avoided are cycles; cyclic definitions may occur, when one resource claims that w_1 is a direct or indirect hyponym of w_2 while another resource contains w_2 as a hyponym of w_1. After the unification of word indices is completed, the resulting hypothesis base is checked for consistency. Some examples of the constraints are given below (x, y stand for word senses, r stands for a relation).

1. **Anti-reflexivity:**
   $\forall x, y, r : r(x, y) \wedge r(y, x) \rightarrow x = y$
2. **Relation uniqueness:**
   $\forall x, y, r_1, r_2 : r_1(x, y) \wedge r_2(x, y) \rightarrow r_1 = r_2$
3. **Transitivity:**
   $\forall x, y, r : r \in Trans \wedge r(x, y) \wedge r(y, x) \rightarrow x = y$

The *anti-reflexivity* constraint claims that terms are not allowed to be connected with themselves. Explicit reflexivity of synonymy is just redundant whereas reflexivity of some other relations (e.g. antonymy, hyponymy, meronymy) is wrong. The *relation uniqueness* constraint claims that only one relation can hold between two word senses. The *transitivity* constraint ensures that for relations that are declared to be transitive (i.e. antisymmetric) no cycles occur.

In our framework, inconsistency is resolved by ranking the axioms provoking the inconsistency by their confidence score. If a relation triple provokes more than one inconsistency then its ranking will be decreased. The relations with the lowest scores are then iteratively excluded until the inconsistency is resolved. If two candidates for exclusion have an equal ranking then the relation triple the removal of which entailes less information loss (checked via inferences) will be eliminated.

Since the WS unification step in our project has not been finished yet, we cannot report about the overall inconsistencies in the integrated structure. However, a preliminary inconsistency evaluation of every single data source is available. For example, 1426 term pairs connected with more than one relation where found in *OpenThesaurus*; *Wiktionary* contains 1696 such pairs; in the *Wortschatz* data no such pairs have been found.

The list of the inconsistency constraints is still open. Probably some more constraints will be identified and added after the first evaluation of the resulting integrated resource has been completed.

5 The LexO architecture

The overall architecture of the LexO framework is displayed in figure 1. We can distinguish three parts: On the lefthand side we find all incoming resources. They provide hypotheses in form of relation triples, which are then integrated by the system. The LexO *engine* (middle) manages the hypothesis database (including confidence values and history for each entry) and its translation to an ontology. On the righthand side we find the output interfaces: A web access and

[7] At present we consider only the most general lexical-semantic relations (synonymy, hyponymy, meronymy, antonymy). If more specific relations will be added, a new methodology of constructing term contexts can turn out to be necessary.

[8] Due to the lack of negation in the relations, the resulting structure cannot become *logically* inconsistent. We rather refer to *structural consistency* here.

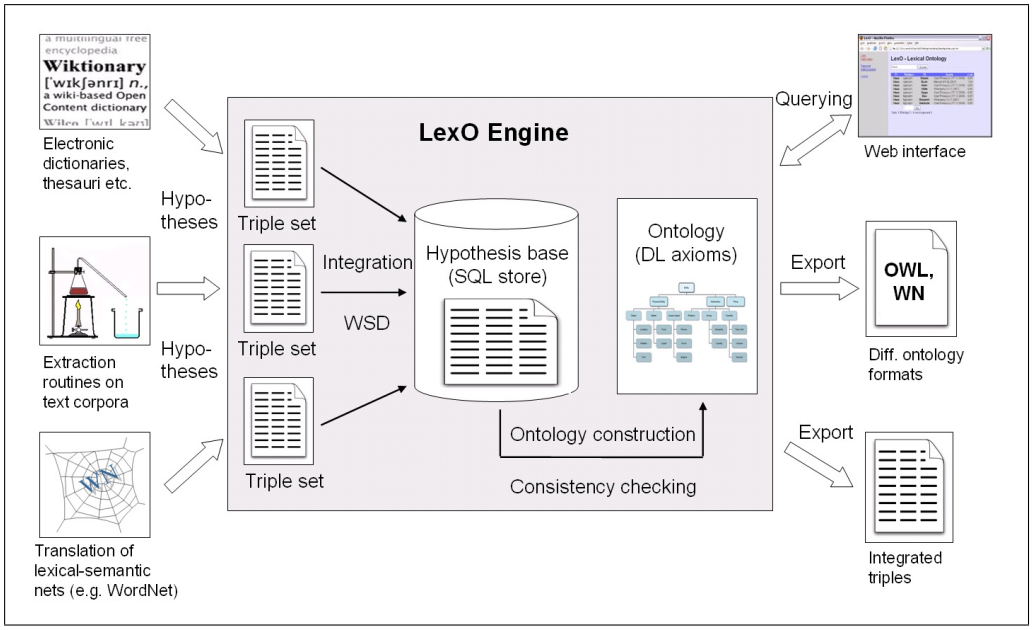

Figure 1: The LexO Architecture

several export routines converting the data to different output formats.

5.1 The hypothesis database

In LexO, a hypothesis is a lexical relation found in some of the various resources that can be used as input. An entry in this database contains the following data:

- The relation itself as a triple `(word-1, relation, word-2)`. If the resource provides sense distinction (as e.g. *Wiktionary* or *OpenThesaurus*), the sense indices are kept together with `word-1` and `word-2`.
- A description of the source of the hypothesis
- A confidence value for the hypothesis (calculated as shown in 3)
- A timestamp indicating when the hypothesis was added

This organisation of the hypothesis base allows for an incremental adding of new hypotheses as well as revision and versioning. For any given point in time the state of the hypothesis base can be reconstructed so that the circumstances leading to a decision in the ontology construction process can be analysed.

5.2 The LexO engine

The LexO engine is the central part of the framework. It manages the database, provides facilities for integrating, filtering and cleaning the raw data (relation triples) and builds up a structured representation assuring pre-defined consistency criteria.

The main problem in the translation process is the confidence-based selection of relations. All data is considered to be more or less reliable (cf. section 3), but, apart from the sanity conditions described in 4.2 (relation uniqueness, connectedness, acyclicity etc.) we have no absolute reliability criterion. We therefore apply a heuristic threshold on the confidence values, depending on the overall growth of the ontology.

5.3 Import-/export interfaces

LexO provides a library of import and export functions as well as a set of interfaces based on it. A number of scripts have been developed to convert each of the resources to triple sets (with *a-priori* confidence values, depending on the resource), and possibly word sense distinction (if provided by the resource). After conversion, a script deals then with the integration of the triples, word sense unification and the import to the triple store (SQL database). In this step, the confidence values are updated, according to the method described in section 3.

The database as well as the ontology can be queried via a web interface (online soon!). This interface will provide masks that allow to search for individual words and relations. Furthermore, another set of converting tools will allow to export the ontology into different formats such as a set of *OWL* clauses or as a WordNet-like database. Methods how to achieve a reasonable *OWL* representation of lexical-semantic relations have been presented by van Assem et al. (2004) and Huang & Zhou (2007). Since plain relations can also be of interest for many applications, a database dump of the hypothesis base will also be provided.

5.4 Evaluation scenarios

Since our project is still in its beginning, we cannot offer any real evaluation of the data yet. However we want to describe here, how an evaluation can be performed. There are basically three complementary strategies: The first is widely used in this domain, because it is straightforward and quick; it is used, for example, by Cimiano et al. (2005). The presupposition is here that we have a reference ontology at hand (gold standard), to which we can compare our data. In the simplest form, we then measure the overlap of relations between our data set and the reference resource in terms of recall and precision. There exist also more complex measures taking the structural similarity into account (cf. Dellschaft & Staab 2006). Our reference resource could be, for example GermaNet, the German word net. However, by determining the overlap of our data with GermaNet, we evaluate obviously not the overall quality, but foremost the

similarity with GermaNet, which is a questionable aspect.

The second strategy relies on direct inspection of the data, it was used, for example, by Snow et al. (2006). Here, human annotators evaluate a representative sample of the constructed data set. Whereas this approach can be very accurate, given the sample is sufficiently large, it implies a lot of efforts and cannot be used for the optimization of confidence parameters (cf. section 3), for example.

The third evaluation scenario is an indirect one. Since the main aim of our project is to serve as a structured semantic resource for NLP tasks, we can evaluate its quality by assessing its performance herein. Harabagiu & Moldovan (2000) for example assess their enriched taxonomy on three tasks: word-sense disambiguation, coreference resolution and information extraction. Measuring the performance of an ontology in such a way implies of course a lot of effort, but it is an objective and independent measure. For this reason we favor this strategy for evaluating the quality of our data.

6 Conclusion and future work

We have proposed an architecture for collecting and integrating lexical-semantic data from various resources. All incoming relations are stored as hypotheses in a database, annotated with automatically determined confidence values. An ontology is created from this hypothesis base by interpreting certain lexical-semantic relations as ontological statements.

We claim that this approach proves especially useful when a broad range of different resources is combined. Therefore we plan to implement additional extraction methods to open up new sources of lexical-semantic information. Beside new sources we will also integrate more types of relations into the database. Apart from that, future efforts will tackle the following issues:

Parameter and threshold estimation: Our project is still in the stage of data cumulation. Whereas we have described in 3, how confidence values can be determined for each relation, we have not optimized the necessary parameters yet. Moreover, we have not yet determined a reasonable threshold for the confidence scores. This kind of parameter tuning takes a lot of time and work and will be subject to our coming efforts.

Creating a common data structure using a top-level ontology: In order to create an ontologically uniform data structure, we want to use a hand-crafted top-level ontology as a seeding ground onto which relations from the database will be successively added. By predefining the top-level concepts we have a means to influence the overall growth of the resulting ontology. However, since the further evolution of the structure strongly depends on the ontological properties of the top-level categorization, it is crucial to construct this structure with a lot of care. Here, the work of Guarino (1998), Gangemi et al. (2002) and Guarino & Welty (2004) will provide valuable guidelines.

Implementation of converters and interfaces: In section 5 we described the output interfaces of our system. These are not implemented yet, but we will try to provide a usable web interface including the possibility to download our data shortly.

Structural constraints and axiomatization: In section 4.2 several simple consistency conditions for our lexical ontology were formulated. However, this set of constraints is definitely not exhaustive. In order to define which constraints are necessary and sufficient for achieving the proposed goals, we need to develop a precise axiomatization of relations and top-level categories in LexO (cf. Gangemi et al. 2001).

7 Acknowledgements

This work is supported by the *Deutsche Forschungsgemeinschaft*, research group "Text Technology" (FOR 437, project C2). The authors also want to thank Helmar Gust, Universität Osnabrück, for fruitful discussions on the subject.

References

Agichtein, E. & Gravano, L. (2000), Snowball: Extracting relations from large plain-text collections, *in* 'Proc. of the 5th ACM International Conference on Digital Libraries (ACM DL)', pp. 85–94.

Atserias, J., Climent, S., Farreres, X., Rigau, G. & Rodriguez, H. (1997), Combining Multiple Methods for the Automatic Construction of Multilingual WordNets, Technical report, Departament de Llenguatges i Sistemes Informatics, Universitat Politecnica de Catalunya, Barcelona.

Brin, S. (1998), Extracting patterns and relations from the world wide web, *in* 'WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98'.

Buitelaar, P., Olejnik, D., Hutanu, M., Schutz, A., Declerck, T. & Sintek, M. (2004), Towards ontology engineering based on linguistic analysis, *in* 'Proc. of the Lexical Resources and Evaluation Conference'.

Buitelaar, P., Olejnik, D. & Sintek, M. (2004), A Protégé plugin for ontology extraction from text based on linguistic analysis, *in* 'Proc. of the 1st European Semantic Web Symposium (ESWS)'.

Buitelaar, P. & Sacaleanu, B. (2001), Ranking and selecting synsets by domain relevance, *in* 'Proc. of NAACL'01'.

Caraballo, S. (1999), Automatic construction of a hypernym-labeled noun hierarchy from text, *in* 'Proc. of the 37th Annual Meeting of the Association for Computational Linguistics', pp. 120–126.

Cederberg, S. & Widdows, D. (2003), Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy, *in* 'Proc. of the Conference on Natural Language Learning'.

Cimiano, P. (2006), *Ontology learning and population from text. Algorithms, Evaluation and Applications*, Springer.

Cimiano, P., A., P., Schmidt-Thieme, L. & Staab, S. (2005), *Learning Taxonomic Relations from Heterogenous Sources of Evidence*, IOS Press, pp. 59–73.

Cimiano, P. & Staab, S. (2004), 'Learning by googling', *SIGKDD Explorations* **6**(2).

Cimiano, P. & Staab, S. (2005), Learning concept hierarchies from text with a guided agglomerative clustering algorithm, *in* 'Proc. of the ICML Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods', Bonn, Germany.

Cimiano, P. & Völker, J. (2005), Text2Onto - a framework for ontology learning and data-driven change discovery, *in* 'Proc.of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB'2005)'.

Cruse, D. A. (1986), *Lexical Semantics*, Cambridge University Press.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. & Harshman, R. (1990), 'Indexing by Latent Semantic Analysis', *JASIS* **41**(6), pp. 391–407.

Dellschaft, K. & Staab, S. (2006), On how to perform a gold standard based evaluation of ontology learning, *in* I. C. et al., ed., 'Proc. of the 5th International Semantic Web Conference (ISWC)', LNCS 4273, Springer Verlag, pp. 228–241.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society* **39**(1), 1–38.

Dorow, B. & Widdows, D. (2003), Discovering corpus-specific word senses, *in* 'Proc. of EACL', Budapest, Hungary., pp. 79–82.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popesu, A.-M., Shaked, T., Soderland, S., S., W. D. & Yates, A. (2004), Web-scale information extraction in KnowItAll, *in* 'Proc. of the 13th World Wide Web Conference', pp. 100–110.

Faure, D. & Nédellec, C. (1998), ASIUM: Learning subcategorization frames and restrictions of selection., *in* 'Proc. of the 10th Conference on Machine Learning (ECML)'.

Fellbaum, C., ed. (1998), *WordNet: An Electronic Lexical Database*, MIT Press.

Gangemi, A., Guarino, N., Masolo, C. & Oltramari, A. (2001), Understanding top-level ontological distinctions, *in* 'Proc. of IJCAI'01'.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. & Schneider, L. (2002), Sweetening ontologies with DOLCE, *in* 'Proc. of EKAW'02'.

Guarino, N. (1998), Some ontological principles for designing upper level lexical resources, *in* 'Proc. of the First International Conference on Lexical Resources and Evaluation', Granada, Spain.

Guarino, N. & Welty, C. (2004), *The Handbook on Ontologies*, Springer-Verlag, chapter An overview of OntoClean, pp. 151–172.

Hamp, B. & Feldweg, H. (1997), Germanet - a lexical-semantic net for german, *in* 'Proc. of the ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications', Madrid.

Harabagiu, S. & Moldovan, D. (2000), *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*, AAAI/MIT Press, chapter Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text, pp. 301–334.

Hearst, M. A. (1992), Automatic acquisition of hyponyms from large text corpora, *in* 'Proc. of the 14th Int. Conf. on Computational Linguistics', Nantes, France.

Huang, X. X. & Zhou, C. L. (2007), 'An OWL-based WordNet lexical ontology', *Journal of Zhejiang University* **8**(6), 864–870.

Kalyanpur, A. (2006), Debugging and Repair of OWL Ontologies, PhD thesis, University of Maryland College Park.

Katrenko, S. & Adriaans, P. (2006), Learning patterns from dependency paths, *in* 'Proc. of the international workhop ontologies in text technology (OTT'06)', Osnabrück.

Khan, L. R. & Hovy, E. H. (1997), Improving the Precision of Lexicon-to-Ontology Alignment Algorithms, *in* 'Proc. of the 1st AMTA Workshop on Interlinguas'.

Knight, K. & Luk, S. K. (1994), Building a Large-Scale Knowledge Base for Machine Translation, *in* 'Proc. of the American Association of Artificial Intelligence AAAI-94.', Seattle, WA., pp. 773–778.

Landauer, T. K. & Dumais, S. T. (1997), 'A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge', *Psychological Review* **104**(1), 211–240.

Maedche, A. (2002), *Ontology learning for the semantic web*, Kluwer.

Naber, D. (2005), OpenThesaurus: ein offenes deutsches Wortnetz, *in* 'Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung 2005.', Peter Lang Verlag, Bonn, pp. 422–433.

Navigli, R. & Velardi, P. (2004), 'Learning domain ontologies from document warehouses and dedicated websites', *Computational Linguistics* **30**(2).

Niles, I. & Pease, A. (2003), Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology, *in* 'Proc. of IKE'03, Las Vegas'.

Okumura, A. & Hovy, E. (1994), Lexicon-to-Ontology Concept Association Using a Bilingual Dictionary, *in* 'Proc. of AMTA, Columbia, MD, 6-8 oct.'.

Oltramari, A., Gangemi, A., Guarino, N. & Masolo, C. (2002), Restructuring WordNet's top-level: The OntoClean approach, *in* 'Proc. of LREC2002 (OntoLex Workshop)'.

Salton, G. & McGill, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

Schütze, H. (1998), 'Automatic word sense discrimination', *Computational Linguistics* **24**(1), pp. 97–124.

Snow, R., Jurafsky, D. & Ng, A. Y. (2006), Semantic taxonomy induction from heterogenous evidence, *in* 'Proc. of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL', Association for Computational Linguistics, Morristown, NJ, USA, pp. 801–808.

van Assem, M., Menken, M., Schreiber, G., Wielemaker, J. & Wielinga, B. (2004), A method for converting thesauri to rdf/owl, *in* 'Proc. of the 3rd Int. Semantic Web Conference (ISWC)', Hiroshima, Japan.

Wagner, A. (2000), Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis, *in* 'Proc. of the ECAI Workshop on Ontology Learning', Berlin, Germany.

Wandmacher, T. (2005), How semantic is Latent Semantic Analysis?, *in* 'Proc. of TALN/RECITAL'05', Dourdan, France.