Mapping Roget's Thesaurus and WordNet to French

Gerard de Melo, Gerhard Weikum

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

Abstract

Roget's Thesaurus and WordNet are very widely used lexical reference works. We describe an automatic mapping procedure that effectively produces French translations of the terms in these two resources. Our approach to the challenging task of disambiguation is based on structural statistics as well as measures of semantic relatedness that are utilized to learn a classification model for associations between entries in the thesaurus and French terms taken from bilingual dictionaries. By building and applying such models, we have produced French versions of Roget's Thesaurus and WordNet with a considerable level of accuracy, which can be used for a variety of different purposes, by humans as well as in computational applications.

1. Introduction

Roget's Thesaurus, first published by Peter Mark Roget in 1852, is surely the most well-known thesaurus in the English-speaking world (Hüllen, 2004), while WordNet (Fellbaum, 1998) is the most widely used lexical database for English natural language processing. In this paper, we describe the techniques we employed to automatically produce translations of the terms in these two remarkable resources from English to French. Our approach relies on translation dictionaries and a set of training mappings to learn a disambiguation model by taking into account statistical properties of the thesaurus and of the dictionary entries (de Melo and Weikum, 2008a). An extension is described that incorporates additional background knowledge from existing thesauri to improve the results.

We are convinced that the resulting resources will facilitate a number of natural language processing and knowledge processing tasks, especially considering the sparsity of freely available alternatives for the French language. Previous studies have used Roget's Thesaurus and WordNet for information retrieval, text classification, word sense disambiguation, thesaurus merging, and semantic relatedness estimation, among other things. Of course, it is also conceivable that the translations be used by authors and editors.

The remainder of this paper is organized as follows. Following a brief introduction of thesauri and related lexical resources in Section 2, we outline the parsing process used to transform Roget's Thesaurus into a machine-readable resource in Section 3. The actual mapping procedure is described in Section 4, beginning with the modelling of the thesaurus and the translation resources, proceeding with the disambiguation model and the feature computation, and finally providing details on an optional extension for using additional background knowledge. Related approaches to the techniques suggested here are referenced in Section 5. Section 6 then provides an evaluation of the resources resulting from our approach, while Section 7 concludes the paper with ideas on future extensions as well as an overall assessment of this work.

2. Thesauri and Similar Resources

The term *thesaurus* may not always be a very clear one because it is commonly used in a variety of contexts to refer to a spectrum of different types of resources. A brief non-exhaustive review could include the following types of thesaurus-like resources:

a) There are terminological resources that conform with official standards such as Z39.19 (ANSI/NISO, 2005) and describe in a well-defined manner the various relations that hold among terms in a specific domain.

b) Such thesauri bear certain similarities with WordNet, a general-purpose lexical resource that explicitly distinguishes different senses of a given term, and provides information about synonymy, hypernymy, and other relationships between such senses (Fellbaum, 1998). An excerpt from WordNet's web interface can be seen in Figure 1, though WordNet is also very commonly used as a machine-readable database for natural language processing.

c) Another more general sense of the term *thesaurus* is used to designate reference works that often provide alphabetical listings of terms with a list of loosely related terms for each headword. Such thesauri tend to be used by writers, although, as mentioned earlier, computational applications have also been explored.

While WordNet is not a thesaurus in this last-mentioned sense, it is used to generate the English thesaurus integrated with the OpenOffice.org office application suite. Roget's Thesaurus, in contrast, indeed can be considered an example of this latter type of thesauri, however with slightly more structural information than many similar thesauri, because the headwords are organized in a complex hierarchy.

3. Parsing Roget's Thesaurus

Since WordNet is a lexical database, accessing the contents is possible in an unambiguous, straight-forward manner. Roget's Thesaurus, despite its astounding level of similarity to more recent resources developed over a hundred and fifty years later, demands additional effort in order to be accessible by current natural language processing tools.

The American 1911 edition of Roget's Thesaurus (Mawson, 1911) has been made available in digital form by Cassidy (2000) with minor extensions, including more than 1,000 new terms and annotations that mark obsolete and archaic forms.

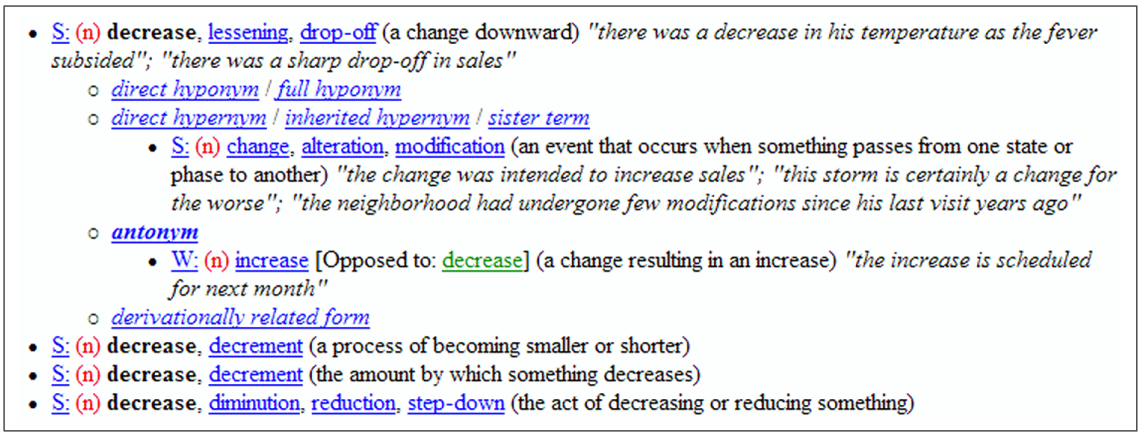

Figure 1: An excerpt from WordNet's web interface, featuring the four noun senses of the term "`decrease`". Additionally, WordNet 3.0 also lists verb senses that were omitted here, and of course a plethora of further related senses is also available on demand.

Although this version is provided as a plaintext file, parsing it in order to obtain a hierarchy of headings and headwords is not quite as trivial as it may seem at first sight. Not only does a myriad of implicit formatting and structuring conventions need to be accounted for, but also the fact that the source file frequently fails to abide to the supposed rules, as there are a considerable number of formatting errors. We used a recursive top-down approach to identify the six top-level classes, which include e.g. "`words relating to the intellectual faculties`", and then proceed to deeper levels. The top-level classes are sometimes subdivided into divisions, e.g. "`communication of ideas`", which consist of sections, e.g. "`modes of communication`". Sections can be further subdivided into multiple levels of subsections, which finally contain headwords. Under each headword one finds one or more part-of-speech markers followed by groups of terms or phrases relating to the headword, as displayed in Figure 2. These groups are delimited by semicolons or full stops, and within such groups, commas or exclamation marks usually fulfil the function of separating individual items, though care needs to be taken not to split up phrases containing such characters. In addition to terms and phrases, these 'semicolon groups' may also contain references to other headwords or to other parts-of-speech of the current headword.

4. The Mapping Process

The mapping from English to French proceeds at the level of basic units, which we simply call thesaurus nodes. In Roget's Thesaurus, we chose the level of semicolon groups rather than the more general headwords in order for the resulting translations to reflect finer distinctions. For instance, in Figure 2, the three terms "`withdraw`", "`take from`", "`take away`" would form a single semicolon group. Within WordNet, we simply regard the individual senses ('synsets') as nodes, e.g. the synset consisting of the terms "`decrease`", "`lessening`", "`drop-off`" with the gloss "`a change downward`" in Figure 1. Our goal is to associate these elementary nodes, which are linked to English terms in the original resources, with the corresponding French terms. For each node, we thus need to determine which of the perhaps many potential translations to use given the English terms of that node and its location in the hierarchy. This disambiguation procedure constitutes the central challenge of the mapping process.

It is quite evident that producing a translation mapping of the lexical units in a thesaurus differs significantly from conventional text translation in several respects. Most importantly, given that the lexical units associated with a node are the focus of our attention, syntactic parsing or syntactic transformations are not required. No attempt is made by our system to translate multi-word expressions or citation phrases that are not in the translation dictionary, because such translations would likely be artificial rather than reflecting a lexicalized, natural use of a term, which is what is expected from a thesaurus. Another major difference lies in the nature of the disambiguation task. Whereas lexical disambiguation in normal machine translation considers the syntactic context in which a particular term occurs, in our case the lexical disambiguation is based on the locus of the node within the hierarchy of the thesaurus.

4.1. Translation Dictionaries

As mentioned earlier, translation dictionaries play a vital role in the mapping process, as they provide the set of candidate translations for each original English term. We used three freely available dictionary packages:

1. the dict-fef software (Dict - Fast - English to French - French to English Dictionary) by Sebastien Bechet, with around 35,000 French-English and around 4,000 English-French translation equivalents (around 22,000 and 2,000 English terms, respectively)
2. the French-English dictionaries from the FreeDict project (Eyermann and Bunk, 2007), originally derived from the Ergane application, with over 16,000 translation equivalents in both directions, covering roughly 9,000 English terms
3. the French-English dictionary from the magic-dic package (Röder, 2002) with around 67,000 translation equivalents for around 20,000 English terms.

```
     #38. Nonaddition. Subtraction. -- N. subtraction, subduction|!;
deduction, retrenchment; removal, withdrawal; ablation, sublation[obs3];
abstraction &c. (taking) 789; garbling,, &c. v. mutilation,
detruncation[obs3]; amputation; abscission, excision, recision; curtailment
&c. 201; minuend, subtrahend; decrease &c. 36; abrasion.
     V. subduct, subtract; deduct, deduce; bate, retrench; remove,
withdraw, take from, take away; detract.
     garble, mutilate, amputate, detruncate[obs3]; cut off, cut away, cut
out; abscind[obs3], excise; pare, thin, prune, decimate; abrade, scrape,
file; geld, castrate; eliminate.
     diminish &c. 36; curtail &c. (shorten) 201; deprive of &c. (take) 789;
weaken.
     Adj. subtracted &c. v.; subtractive.
     Adv. in deduction &c. n.; less; short of; minus, without, except,
except for, excepting, with the exception of, barring, save, exclusive of,
save and except, with a reservation; not counting, if one doesn't count.
```

Figure 2: Excerpt from Roget's Thesaurus text file.

These dictionaries were parsed in order to convert the data and annotations into a machine-processable format. Since the first two packages offer separate English-French as well French-English mappings, we effectively coalesced entries from five dictionaries into a unified translation knowledge base with around 78,000 translation equivalents and coverage of around 34,000 English terms and 48,000 French terms.

From the dictionaries we imported not only the raw mappings from terms in one language to terms in another language, but also restrictions on the part-of-speech for the source or target term. For instance, for the translation from English "`store`" to French "`enregistrer`", the dictionary additionally tells us that this translation applies to the word "`store`" as a verb. For the coalescing, we adopted the policy of letting entries with part-of-speech information override entries without such information.

4.2. Disambiguation Model

Translation dictionaries often provide many different translations for an English term, but typically only few of these translations are appropriate for a given node in the thesaurus. For instance, the term "`store`", depending on the particular thesaurus node, could be translated as "`approvisionner`", "`entrepôt`", "`boutique`", "`emmagasiner`", "`bouillon`", or "`magasin`", among others.

Our disambiguation approach is an application of a technique we initially developed to generate a German-language version of WordNet (de Melo and Weikum, 2008a) that has now been extended to include several novel statistics. The basic idea is to use supervised machine learning to derive a model for classifying translations from manually labelled translation pairs. We conceive each semicolon group (or synset) in the thesaurus as a separate node to be translated to one or more terms. Since a good coverage of the target language is an important desideratum, we allow for translating a single English term to multiple French terms whenever this is appropriate. Furthermore, nodes may also remain vacuous when no adequate translation is available, as many thesauri are designed to cover a wide range of terms, including rare and obsolete terms that may often be untranslatable. This is most certainly the case for Roget's Thesaurus, bearing in mind that merely 41% of the terms in the 1987 Penguin Edition are covered by WordNet 1.6 according to Jarmasz and Szpakowicz (2001).

Given a French target term t and a thesaurus node n, we considered the tuple (n, t) a candidate mapping if and only if one of the English source terms associated with n in the original thesaurus is translated as t according to the unified translation knowledge base. Such tuples can either represent appropriate translations (positive examples) or inappropriate ones (negative). To produce a training data set, 731 training such candidate pairs were manually evaluated for Roget's Thesaurus (31% positive), and, likewise, 611 such training mappings (33% positive) were established for WordNet. For each node-term pair, we created a real-valued feature vector in an m-dimensional Euclidean space $\mathbb{R}^m$, as will further be elucidated later on. The training feature vectors were then used to derive a model using a linear kernel support vector machine (SVM) (Vapnik, 1995). The model, computed with LIBSVM (Chang and Lin, 2001), makes a prediction about whether a new feature vector in this Euclidean space is more likely to be a positive or a negative instance. In effect, it induces a decision rule for whether a given French term t should serve as one of the translations for a node n. Applying the model to feature vectors representations of all candidate mappings (n, t), we were able to decide which translations to accept to produce the output, i.e. an association of French terms to the nodes in the thesaurus hierarchy. Rather than being a direct acknowledgment of the support vector model's decision hypersurface, our decision rule is based on first estimating posterior probabilities $p_{n,t}$ from the SVM outputs (Platt, 1999), and then using two thresholds $p_{\min}$ and $p_{\min*} \leq p_{\min}$, where a pair (n, t) is accepted if either $p_{n,t} > p_{\min}$, or alternatively $p_{n,t} > p_{\min*}$ and $\neg\exists t' : p_{n,t'} > p_{n,t}$.

4.3. Feature Computation

We now turn our attention to the feature values $x_0, \ldots, x_{m-1}$ that make up the m-dimensional real-valued feature vector $\vec{x} \in \mathbb{R}^m$ for a given node-term pair (n, t). Each feature value x_i is a score that attempts to discern and quantify some information about the pair, based on the thesaurus, the unified translation knowledge base, or additional external knowledge sources.

Some of the most significant scores assess the similarity of each of the translations e of the term t to the current node n using the following formula:

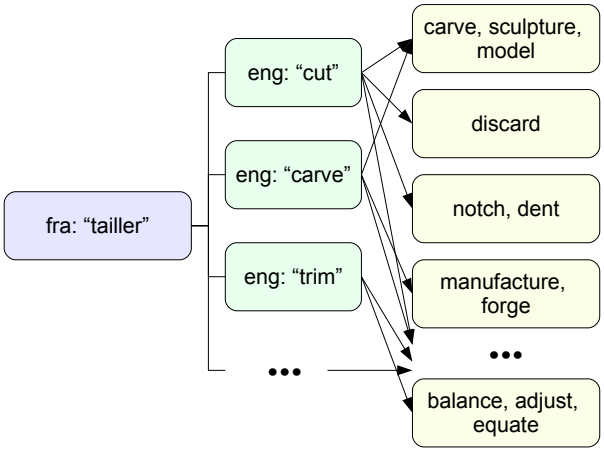

Figure 3: Links from a French term via English translations to nodes in Roget's Thesaurus (labels abbreviated). Due to the polysemy of the English terms, many nodes are inappropriate for the French term.

$$\sum_{e \in \phi(t)} \max_{n' \in \sigma(e)} \gamma(t, e, n') \, \mathrm{sim}_{\mathrm{n}}(n, n') \tag{1}$$

It is assumed here that $\phi(t)$ yields the set of English translations of a French term t, $\sigma(e)$ yields the set of thesaurus nodes of an English term e, $\gamma(t, e, n)$ is a weighting function for term-node pairs n, t with respect to a translation e, and $\mathrm{sim}_{\mathrm{n}}(n_1, n_2)$ is a semantic relatedness measure between two thesaurus nodes n_1 and n_2. In Figure 3, we observe that the uppermost node is relevant for both “`cut`” and “`carve`” by being directly linked to them. Using advanced relatedness measures $\mathrm{sim}_{\mathrm{n}}$ we also can account for cases where there is no such direct link. For Roget's Thesaurus, apart from using the trivial identity test, we also computed such scores using a graph distance similarity measure. For WordNet, we used the identity test, the graph distance algorithm, as well as a gloss similarity measure, as described in more detail elsewhere (de Melo and Weikum, 2008a).

The weighting function scores $\gamma(t, e, n)$ are mainly used to pay respect to the extent of the relevance of other nodes n' for our current term t. For example, we may disregard nodes that are known or predicted to have a different part-of-speech than the term t using a lexical category compatibility weighting function, or we may weight WordNet nodes by their sense frequency in the SemCor corpus (de Melo and Weikum, 2008a).

Several further scores attempt to quantify the number of relevant alternatives to n when considering thesaurus nodes for t. A global score is computed as follows:

$$\frac{1}{\sum_{n' \in C} (1 - \mathrm{sim}_{\mathrm{n}}(n, n'))} \tag{2}$$

where $C = \bigcup_{e \in \phi(t)} \sigma(e)$ stands for the complete candidate set. The score assesses how many relevant alternative nodes there are, or, metaphorically speaking, how many rivals there are, because a lower number of rivals increases our confidence in the acceptance of (n, t).

Additionally, one can also consider what might be called a local variant of the score. For each English translation e of t, we may distinguish three cases: (1) e is directly connected to n (2) e is not directly connected to n but instead to some other n' that is sufficiently similar to n (3) none of the nodes n' that e is linked to exhibit a sufficient level of resemblance with n. For cases (1) and (2) we may then ask how many relevant alternative nodes there are, and hence introduce several scores of the following form:

$$\sum_{e \in \phi(t)} \frac{\mathbf{1}_{\{n_1 \mid \exists n_2 \in \sigma(e):\, \mathrm{sim}_{\mathrm{n}}(n_1, n_2) \geq s_{\min}\}}(n)}{1 + \sum_{n' \in \sigma(e)} \gamma(t, e, n')(1 - \mathrm{sim}_{\mathrm{n}}(n, n'))} \tag{3}$$

Here, $\mathbf{1}_S$ is the indicator function for a set S ($\mathbf{1}_S(s) = 1$ if $s \in S$ and 0 otherwise). We use it to take into account only those translations e that are actually linked to n or to some node n' with a high similarity to n (note, however, that we set the similarity threshold $s_{\min}$ to 1.0).

Similarly, we observed that the number of alternative French terms t that could be associated with a given thesaurus node n can serve as an indication of whether to accept a mapping. In the extreme case, when no other French terms other than our current term t are eligible for the current node, the chances of a positive match are rather high. This gives rise to the following formula, which is symmetric to Equation 3 above:

$$\sum_{e \in \sigma^{-1}(n)} \frac{\mathbf{1}_{\{t_1 \mid \exists t_2 \in \phi^{-1}(e):\, \mathrm{sim}_{\mathrm{t}}(t_1, t_2) \geq s_{\min}\}}(t)}{1 + \sum_{t' \in \phi^{-1}(e)} \gamma(t', e, n)(1 - \mathrm{sim}_{\mathrm{t}}(t, t'))}. \tag{4}$$

Here, $\sigma^{-1}(n) = \{e \mid n \in \sigma(e)\}$ yields the set of all English terms associated with a node, $\phi^{-1}(e) = \{t \mid e \in \phi(t)\}$ yields the set of all French translations of an English term, and $\mathrm{sim}_{\mathrm{t}}(t_1, t_2)$ is a semantic relatedness function between French terms, in our case either a simple identity test, or a more advanced measure as described later in Section 4.4.

Apart from these scores, we additionally integrate the scores that had previously been used for building a German-language WordNet (de Melo and Weikum, 2008a). Further improvements were obtained by considering scores that capture the relative placement of a node with reference to the other nodes under consideration for a term t. Given a feature score $f(n, t)$, say a semantic overlap score as described by Equation 1, as well as a minimal score value $f_{\min}$ that would ideally be attained by at least one of the nodes, we calculate a corresponding relative score

$$\frac{f(n, t)}{\max(\{f_{\min}\} \cup \{f(n', t) \mid n' \in C\})} \tag{5}$$

with respect to the set of candidate nodes C as defined above.

4.4. Additional Background Knowledge

After computing the training feature vectors in this way, we used the learnt model to obtain probabilities $p_{n,t}$ for all potential candidate pairs (n, t) as described earlier in Section 4.2. For Roget's Thesaurus, we then also evaluated

```
réduction|1
(Nom)|abrègement|diminution|troncation|siglaison|assouplissement|
modération|mesure|atrophie|maigreur|cachexie|compactage|densification|
entassement|amoindrissement|soustraction|décroissance|retranchement|
abaissement|dévalorisation|dévaluation|amenuisement|amincissement|
amaigrissement|dépréciation|discount|remise|escompte|limitation|borne|
bornage|restriction|délimitation|démarcation|maquette|ébauche|modèle|
copie|projet|canevas|pondération|rationnement|régime|pacification|
ristourne|simplification
réduire|1
(Verbe)|abréger|résumer|raccourcir|condenser|écourter|restreindre|
diminuer|rapetisser|amoindrir|rétrécir|atrophier|amaigrir|atténuer|
affaiblir|alléger|tempérer|adoucir|soulager|minimiser|excuser|
compacter|compresser|serrer|rogner|couper|tronquer|tasser|laminer|
aplatir|écraser|étirer|user|limiter|borner|circonscrire|délimiter|
démarquer|contingenter|arrêter|localiser|plafonner|entourer|cerner|
optimiser|améliorer|maximaliser|perfectionner|culminer|préciser|
définir|énoncer|établir|expliciter|détailler|clarifier|éclairer|
souligner|fixer|spécifier|caractériser|ralentir|freiner|modérer|
retarder|ramener|amener|rétablir|raréfier|rationner|mesurer|répartir|
soumettre|cantonner|simplifier|schématiser|subjuguer|asservir|dominer|
conquérir|dompter|charmer|enchanter|envoûter|capter|captiver
```

Figure 4: Excerpt from French thesaurus available in conjunction with the OpenOffice.org 2.0 office application suite

an optional supplementary coverage boosting procedure, based on the availability of additional prior knowledge. We parsed the French thesaurus from the OpenOffice.org 2.0 application suite[1] to gain relationship information between French terms (cf. Figure 4). The feature computation and classification process was then re-iterated using several new extended scores that exploit this additional knowledge as well as the initial, preliminary probabilities $p_{n,t}$.

First of all, the information from the external resource allows us to define a similarity measure sim_t between French terms, where $\mathrm{sim}_\mathrm{t}(t,t') = 1$ if the two terms are directly related according to the OpenOffice.org thesaurus, and 0 otherwise. We may then define $R(t) = \{t' \mid \mathrm{sim}_\mathrm{t}(t,t') = 1\}$ as the neighbourhood of t, and compute a score of the following form

$$\sum_{t' \in R(t)} \max_{e \in \phi(t')} \max_{n' \in \sigma(e)} \gamma(t', e, n')\, \mathrm{sim}_\mathrm{n}(n, n') \qquad (6)$$

This formula assesses the similarity of each related term t' to the candidate node n by computing the maximum similarity to n of any of the senses indirectly linked to t' via translations. Figure 5 shows how these related terms may reinforce candidate nodes. The choice of a weighting function $\gamma(t,e,n) = p_{n,t}$ based on the initial probabilities leads to a score that essentially reflects whether any of the related terms are also being mapped to the current thesaurus node. For instance, even if a term t has a low initial probability $p_{n,t}$, the fact that it is known to be related to several other terms t' with high probabilities $p_{n,t'}$ may increase the prospects of (n,t) being accepted with little risk of committing an error. Moreover, mutual reinforcement of multiple pairs (n,t) with low probabilities is also possible.

The relatedness measure between French terms can further be applied to Equation 4, again in conjunction with a weighting function $\gamma(t,e,n) = p_{n,t}$ that uses the initial classification probabilities. The equation then reflects the number of alternative French terms that could be associated with the node, weighted by their initial probability estimate, and disregarding those alternative terms that are known to be related to the current term t under consideration.

[1] http://www.openoffice.org/

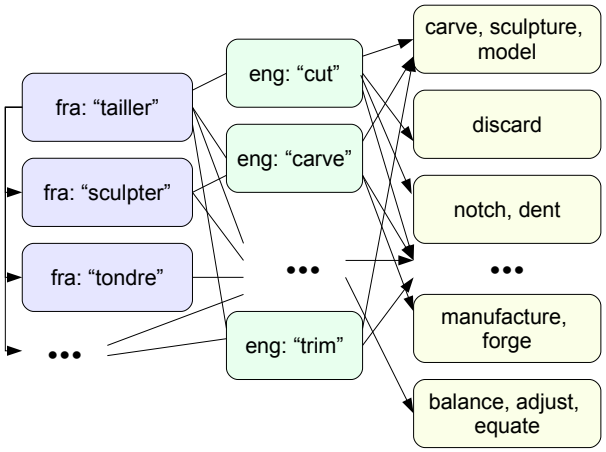

Figure 5: Indirect connections from a French term to thesaurus nodes via related terms (additional background knowledge) and translations.

Finally, we can use the new similarity measure to compute a reverse form of Equation 1:

$$\sum_{e \in \sigma^{-1}(n)} \max_{t' \in \phi^{-1}(e)} \gamma(t', e, n')\, \mathrm{sim}_\mathrm{t}(t, t') \qquad (7)$$

Here, one considers all the English terms associated with a thesaurus node, retrieves their translations, and then determines to what extent these translations are similar to the current term t under consideration. Certainly, this could also be computed with a trivial identity similarity measure for French terms, but due to the symmetric behaviour of the translation functions ϕ and ϕ^{-1}, the resulting values would correspond to those computed using Equation 1.

With these new feature scores, we recreated all feature vectors and trained a new classification model.

5. Related Work

Scannell (2003) describes an approach that seems to be grounded in similar intuitions as ours, which he used to connect Irish Gaelic terms with English terms in Roget's

Thesaurus. In contrast to our system, the disambiguation is based on a simple heuristic combined with manual disambiguation work, and the final thesaurus does not retain the original structure of Roget's Thesaurus.

A number of works have considered mapping non-English terms to Princeton WordNet, e.g. Okumura and Hovy (1994), Atserias et al. (1997), and Sathapornrungkij and Pluempitiwiriyawej (2005), adopting what Vossen (1996) has called the *expand model* for adapting WordNet to new languages. The overall strategy adopted in these studies is comparable to ours in that translation dictionaries are used to translate an existing lexical resource based on disambiguation heuristics, however they rely on static predetermined criteria that are often WordNet-specific rather than flexibly adapting to the mapping task by learning a decision rule from real-valued feature vector instances.

Other researchers have proposed deriving thesaurus-like information from corpus co-occurrence statistics (Doyle, 1961; Grefenstette, 1993; Jacquemin and Ploux, 2006) or from monolingual dictionaries (Ploux and Victorri, 1998), however the resulting resources lack the well-organized hierarchy and semantic links of Roget's Thesaurus and WordNet. Although there are, in principle, means of detecting hypernym relations from text corpora (Hearst, 1992), such techniques tend to have a low recall so only a small subset of all terms will end up being linked, and it is unlikely that a coherent organization scheme could be constructed from such individual links. A more feasible alternative is to enrich an existing thesaurus with new additional information derived from corpora or dictionaries (Araujo and Pérez-Agüera, 2006). Such techniques could also be invoked to enhance the thesauri generated by our mapping procedure.

6. Evaluation

In this section, we report on our evaluation of the French thesaurus translation mappings. Using the same procedure as for the training data sets, we generated test sets that are independent and do not overlap in any way with the training data. Our test set for Roget's consists of 1,012 human-evaluated n, t pairs, labelled as either positive or negative examples, while the respective WordNet mappings numbered 1,276. As is common in classification settings, we evaluate the classification model using the accuracy, precision, and recall as evaluation measures.

6.1. Roget's Thesaurus

As mentioned earlier, we first built a version of Roget's Thesaurus without additional background information from the OpenOffice.org thesaurus. Table 1 shows the precision and recall values on the test set for this version using several choices of $p_{\min}$ and $p_{\min*}$, thereby demonstrating the trade-off between precision and recall.

The results are quite impressive, given the enormous difficulty of this task, and demonstrate the viability of our approach despite our designation of semicolon groups as the basic units, which requires much finer distinctions than would be necessary at the level of headwords.

Table 1: Comparison of precision and recall of Roget's Thesaurus translation for different choices of classification thresholds, based on 1,012 manually created test mappings

$p_{\min}$	$p_{\min*}$	precision	recall
0.3	0.25	84.05%	77.75%
0.35	0.3	85.80%	75.50%
0.5	0.5	89.49%	66.00%
0.6	0.5	89.38%	61.00%

Table 2: Comparison of precision and recall of Roget's Thesaurus translation (with additional background information from the OpenOffice.org thesaurus) for different choices of classification thresholds, based on 1,012 manually created test mappings

$p_{\min}$	$p_{\min*}$	precision	recall
0.3	0.25	84.94%	81.75%
0.35	0.3	87.64%	78.00%
0.5	0.5	89.40%	67.50%
0.6	0.5	91.01%	63.25%

6.2. Roget's Thesaurus with Background Knowledge

Subsequently, we evaluated how background knowledge from the OpenOffice.org thesaurus can be leveraged to improve the quality of the mappings using the scheme described in Section 4.4.

The results are shown in Table 2. Having taken a closer look at the generated output, we noticed that many of the misclassifications at the semicolon level are nevertheless correct when considered at the level of head words, so we opted for $p_{\min} = 0.3$, $p_{\min*} = 0.25$ for the rest of our study, which produces a more extensive coverage than higher thresholds. It is quite obvious that other choices may be embraced when the prospective applications require fine-grained distinctions with a greater level of accuracy. Since most modern natural language processing applications rely on some form of statistical processing, another alternative would be immediately adopting the thesaurus as a probabilistic one, where every term is considered linked to a node with a certain probability.

With our choice of parameter values for $p_{\min}$ and $p_{\min*}$, we generated the final version of Roget's Thesaurus, and evaluated it in more detail as shown in Tables 3 and 4. The former reports the precision and recall by part-of-speech, showing how verbs are much harder to classify correctly due to their greater polysemy, while the latter table offers coverage statistics. Figure 6 shows an excerpt from the generated thesaurus.

6.3. WordNet

The same classification approach was also applied to WordNet, and results on the test set for several acceptance thresholds are shown in Table 5. The overall scores imply slightly less accurate results than for Roget's Thesaurus, which, of course, is due to the fact that the sense distinctions are even more subtle than for Roget's semicolon groups. Indeed, even humans often have considerable difficulties when annotating text with WordNet senses. Additional preliminary

```
    #38. N. soustraction; ratiocination; abaissement, retranchement, réduction;
ablation, retrait; abstraction, idée abstraite; mutilation; amputation;
abaissement, raccourcissement, réduction, diminution; décroîs, abaissement;
abrasion;
    V. prélever, soustraire, déduire, retrancher; déduire; retirer, emporter,
ôter, éloigner, restreindre, claustrer; éloigner, ôter, prélever, emporter,
retrancher; fausser, mutiler, amputer; retrancher, abattre, tailler; découper;
rogner, diluer, délayer, exciser, éclaicir; décimer, racler; limer; châtrer,
castrer; supprimer; rapetisser; écourter, restreindre, raccoucir, abréger;
    Adj. soustrait;
    Adv. en outre, excepté, moins, épargnons, dénué de, sans, sauvons, sauf;
```

Figure 6: Excerpt from translation of Roget's Thesaurus text file. Note how polysemy can lead to mistranslations (translating the English "`deduction`" to "`ratiocination`" may make sense in certain contexts, however in this case a different sense of "`deduction`" was intended).

Table 3: Evaluation of mapping classifications for Roget's Thesaurus with additional background knowledge using $p_{\min} = 0.3$, $p_{\min*} = 0.25$ and based on 1,012 manually created test mappings

	precision	recall
nouns	83.98%	88.27%
verbs	76.56%	60.49%
adjectives	91.75%	88.12%
adverbs	88.89%	72.73%
overall	84.94%	81.75%

Table 4: Coverage statistics for translation of Roget's Thesaurus with additional background knowledge using $p_{\min} = 0.3$, $p_{\min*} = 0.25$

	terms	lexicalized nodes	node mappings
nouns	11,161	11,628	31,376
verbs	3,624	4,861	14,666
adjectives	6,166	5,418	15,116
adverbs	705	651	1,638
total	21,232	22,560	62,798

tests also indicated that the OpenOffice.org thesaurus data was not well-suited for producing significant improvements to the WordNet mappings. Comparing Figures 1 and 4, it should be quite clear that the OpenOffice.org thesaurus information is often too coarse.

Though lower than for Roget's Thesaurus, the overall results fulfil the demands of many applications, in particular for human thesauri where fine-grained sense distinctions are not a requirement, and for statistical natural language processing systems that are able to benefit from imperfect mappings (de Melo and Weikum, 2008b). Using the settings $p_{\min} = 0.7$ and $p_{\min*} = 0.6$, a more in-depth assessment was carried out, as shown in Tables 6 and 7.

Obviously, automatically translated resources may not be able to deliver the same wealth of useful terms and expressions and the same level of organization as a native thesaurus compiled over many years by lexicographers and native speakers familiar with a wide range of literature. For WordNet, this difference is more pronounced than for Roget's Thesaurus because an ideal French WordNet would have French-language glosses and additional synsets to cover French-specific terms (de Melo and Weikum, 2008b).

Table 5: Comparison of precision and recall of WordNet translation for different choices of classification thresholds, based on 1,276 manually created test mappings

$p_{\min}$	$p_{\min*}$	precision	recall
0.5	0.5	72.78%	76.88%
0.6	0.55	78.00%	73.13%
0.7	0.6	81.20%	63.44%
0.8	0.7	86.22%	52.81%

Table 6: Evaluation of mapping classifications for WordNet, using $p_{\min} = 0.7$ and $p_{\min*} = 0.6$ and based on 1,276 manually created test mappings

	precision	recall
nouns	80.51%	63.76%
verbs	59.46%	47.83%
adjectives	91.67%	69.47%
adverbs	86.96%	66.67%
overall	81.20%	63.44%

However, the coverage of these WordNet mappings is quite encouraging and could be improved even more by exploiting more complete translation dictionaries.

7. Concluding Remarks

In this paper, we described a framework that successfully allowed us to automatically create French versions of Roget's Thesaurus and WordNet, based on freely available translation dictionaries and a set of manually established training mappings. Sense disambiguation is performed by training and applying a linear model that, given a possible mapping from an entry or node in the thesaurus to a French term, evaluates various structural and statistical properties

Table 7: Coverage statistics for translation of Princeton WordNet, using $p_{\min} = 0.7$ and $p_{\min*} = 0.6$

	terms	lexicalized synsets	sense mappings
nouns	14,563	15,565	25,141
verbs	3,807	4,765	9,717
adjectives	6,637	6,548	12,621
adverbs	1,234	1,252	2,069
total	25,716	28,130	49,548

of the translations and of the connected thesaurus nodes to predict whether the mapping should be accepted.

Several extensions to the work described in this paper merit further consideration in the future. We are currently investigating possible ways to consolidate the lexical knowledge from the two generated resources as well as from the OpenOffice.org thesaurus into a single, more general and comprehensive lexical resource. Research is necessary on how the different organizational paradigms can be meshed in a sensible way. Another area of research is the use of corpora to validate the information in the translated thesauri as well as to augment them with further terms, including perhaps domain-specific vocabulary. Section 4.4 already provides a framework that could be used to incorporate corpus-derived information, however additional experiments are necessary in other to evaluate whether adjustments need to be made. Finally, we would like to study how the two resources we have generated can best be exploited in information retrieval settings.

Since the French component of the EuroWordNet project is not freely available, we believe that these two resources will be very useful for natural language processing tasks and for human users. Apart from the obvious benefits of integration with word processors, these resources may also be used for computational tasks such as query expansion, semantic relatedness estimation, or cross-lingual information retrieval.

8. References

ANSI/NISO. 2005. *Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. NISO Press, Bethesda, MD, USA.

Lourdes Araujo and José R. Pérez-Agüera. 2006. Enriching thesauri with hierarchical relationships by pattern matching in dictionaries. In *Proc. 5th International Conference on NLP FinTAL*, pages 268–279.

Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodríguez. 1997. Combining multiple methods for the automatic construction of multilingual WordNets. In *Proc. Conf. Recent Advances in NLP 1997*, pages 143–149.

Patrick Cassidy. 2000. An investigation of the semantic relations in the roget's thesaurus: Preliminary results. In *Proc. CICLing 2000, International Conference on Intelligent Text Processing and Computational Linguistics*, pages 181–204.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

Gerard de Melo and Gerhard Weikum. 2008a. A machine learning approach to building aligned wordnets. In *Proceedings of the First International Conference on Global Interoperability for Language Resources*.

Gerard de Melo and Gerhard Weikum. 2008b. On the utility of automatically generated wordnets. In *Proc. Global WordNet Conference*.

Lauren B. Doyle. 1961. Semantic road maps for literature searchers. *J. ACM*, 8(4):553–578.

Horst Eyermann and Michael Bunk, 2007. *FreeDict*. `http://www.freedict.org/en/`.

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

Gregory Grefenstette. 1993. Automatic thesaurus generation from raw text using knowledge-poor techniques. In *Making Sense of Words. 9th Annual Conference of the UW Centre for the New OED and Text Research*.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational Linguistics*, pages 539–545, Morristown, NJ, USA. Association for Computational Linguistics.

Werner Hüllen. 2004. *A History of Roget's Thesaurus. Origins, Development, and Design*. Oxford University Press.

Bernard Jacquemin and Sabine Ploux. 2006. Corpus spécialisé et ressource de spécialité : l'information forme le sens. In *Journées Scientifiques du CRTT : Corpus et dictionnaires de langues de spécialité*.

Mario Jarmasz and Stan Szpakowicz. 2001. The design and implementation of an electronic lexical knowledge base. In *Proc. 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (AI 2001)*, pages 325–333.

C.O. Sylvester Mawson, editor. 1911. *Roget's Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition*. McDevitt-Wilson's, Inc., New York, NY, USA.

Akitoshi Okumura and Eduard Hovy. 1994. Building Japanese-English dictionary based on ontology for machine translation. In *Proc. Workshop on Human Language Technology*, pages 141–146.

John C. Platt, 1999. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA.

Sabine Ploux and Bernard Victorri. 1998. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement automatique des languages*, 39:161–182.

Jens Röder. 2002. The magic-dictionary magic-dic.

P. Sathapornrungkij and C. Pluempitiwiriyawej. 2005. Construction of Thai WordNet lexical database from machine readable dictionaries. In *Proc. 10th Machine Translation Summit, Phuket, Thailand*.

Kevin Scannell. 2003. Automatic thesaurus generation for minority languages: an Irish example. In *Actes de la 10e conférence TALN, volume 2, Workshop Traitement Automatique des Langues Minoritaires et des Petites Langues*, pages 203–212.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Piek Vossen. 1996. Right or wrong: Combining lexical resources in the EuroWordNet project. In *Proc. Euralex-96*, pages 715–728.