Cross-lingual porting of distributional semantic classification

Lilja Øvrelid
Department of Linguistics
University of Potsdam
Germany
lilja@ling.uni-potsdam.de

Abstract

This article presents experiments in the porting of semantic classification between two closely related languages, Swedish and Danish. We show that a classifier for the semantic property of animacy, trained on morphosyntactic distributional data for one language may be applied directly to data from another language with little loss in terms of accuracy.

1 Introduction

Semantic classification of natural language has in recent years received extensive attention.[1] Most approaches to these tasks make use of language-specific, annotated data or lexical resources, such as FrameNet and WordNet, a fact which complicates a multilingual perspective on semantic annotation and classification. One way of approaching this is found in work on projection of semantic classifications, such as semantic roles, making use of parallel corpora and hence the relation of translation to acquire semantic relations for new languages (Pado and Lapata, 2005; Johansson and Nugues, 2006).

Much recent work in semantic classification assumes that the syntactic distribution of lexical items constitutes a reliable predictor of semantics or meaning, at the *type* level (Lin, 1998). In the task of verb classification, for instance, it has been shown that features motivated in typological generalizations and found to be highly predictive for classification in one language (English) may be 're-used' for the classification of verbs in other languages, such as Italian (Merlo et al., 2002). The semantic property of animacy influences linguistic phenomena in a range of different languages, and has been shown to correlate quite reliably with other semantic, syntactic and information-structural properties, such as agentivity, argumenthood and topicality (de Swart et al., 2008). In computational linguistic work, animacy has been shown to provide important information in anaphora resolution (Orăsan and Evans, 2007), argument disambiguation (Dell'Orletta et al., 2005) and syntactic parsing in general (Øvrelid and Nivre, 2007).

In this article we will explore the porting of a semantic classifier from one language to another, investigating the application of a semantic classifier trained on distributional data for one language directly to data from another language. We present first experiments examining the porting of automatic classification for the semantic property of animacy between the closely related languages of Swedish and Danish. Unlike previous work, we do not assume a parallel corpus or a gold standard annotation for the second language (Danish).

2 Swedish animacy classification

Talbanken05 is a Swedish treebank converted to dependency format, containing both written and spoken language (Nivre et al., 2006b).[2] In addition to information on part-of-speech, dependency head and relation, Talbanken05 distinguishes animacy for all nominal constituents.[3]

The dimension of animacy roughly distinguishes between entities which are alive and entities which are not. Table 1 presents an overview

[1]Parts of the research reported in this paper has been supported by the *Deutsche Forschungsgemeinschaft* (DFG, *Sonderforschungsbereich 632*, project D4).

[2]The written sections of the treebank consist of professional prose and student essays and amount to 197,123 running tokens, spread over 11,431 sentences.

[3]To be precise, the annotation in Talbanken05 distinguishes between 'person' and 'non-person'.

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 246–249

Class	Types	Tokens covered
Animate	644	6010
Inanimate	6910	34822
Total	7554	40832

Table 1: The animacy data set from Talbanken05; number of noun lemmas (Types) and tokens in each class.

of the animacy data for common nouns in Talbanken05. It is clear that the data is highly skewed towards the 'inanimate' class, which accounts for 91.5% of the data instances. Due to the small size of the treebank we classify common noun *lemmas*. Following a strategy in line with work on verb classification (Merlo and Stevenson, 2001; Stevenson and Joanis, 2003), we set out to classify the lemmas based on their morphosyntactic distribution in a considerably larger corpus. For the animacy classification of common nouns, we construct a general *feature space* for animacy classification, which makes use of distributional data regarding syntactic properties of the noun, as well as various morphological properties. The syntactic and morphological features are presented below:

Syntactic features subject (SUBJ), object (OBJ), prepositional complement (PA), root (ROOT), apposition (APP), conjunct (CC), determiner (DET), predicative (PRD), complement of comparative subjunction (UK).

Morphological features gender (NEU/UTR), number (SIN/PLU), definiteness (DEF/IND), case (NOM/GEN).

For each noun lemma w, relative frequencies of the morphosyntactic features f_i are calculated from the corpus: $\frac{freq(f_i,w)}{freq(w)}$. For extraction of distributional data for the Talbanken05 nouns we make use of the Swedish Parole corpus of 21.5M tokens,[4] and to facilitate feature extraction, we part-of-speech tag the corpus and parse it with MaltParser[5] (Nivre et al., 2006a), which assigns a dependency analysis.[6]

[4]Parole is available at http://spraakbanken.gu.se

[5]http://www.maltparser.org

[6]For part-of-speech tagging, we employ the MaltTagger – a HMM part-of-speech tagger for Swedish (Hall, 2003). For parsing, we employ MaltParser with a pretrained model for Swedish, which has been trained on the tags output by the tagger.

Classification is performed with Support Vector Machines (SVMs) and we make use of the LIBSVM package (Chang and Lin, 2001) with a RBF kernel ($C = 8.0, \gamma = 0.5$).[7] For training and testing of the classifiers, we make use of leave-one-out cross-validation.

We obtain results for animacy classification, ranging from 97.1 accuracy to 93.7 depending on the sparsity of the data.[8] With an absolute frequency threshold of 10, we obtain an accuracy of 95.1%, which constitutes a 46.7% reduction of error rate compared to a majority baseline which assigns the class of inanimate to all instances (90.8).

3 Danish distributional data

The Swedish classifier has been trained on distributional data which generalizes over the distribution of individual nouns. In order to apply the animacy classifier trained on Swedish and described in section 2 above, we will need morphosyntactic distributional data for Danish noun lemmas along the same set of features as those employed for the classification of Swedish nouns.

3.1 Data

We employ the freely available Danish corpus Korpus2000 which contains approximately 22 million words.[9] In order to obtain both morphological and syntactic information regarding the nouns in the corpus, we part-of-speech tag and parse the corpus, employing MaltTagger and MaltParser, both trained on the analysis found in the Danish Dependency Treebank (DDT) (Kromann, 2003).[10].

3.2 Features

For application of the animacy classifier to Danish data, we must also represent our data set within the same feature space as the one defined for the Swedish classification task. As mentioned earlier, Korpus2000 has been parsed with a parser which assigns a dependency analysis, and followingly has much in common with the dependency analysis in Talbanken05. Even so, the syntactic anal-

[7]Parameter optimization, i.e., choice of kernel function, C and γ values, is performed on 20% of the total data set with the `easy.py` tool, supplied with LIBSVM.

[8]With a threshold of 1000 instances in Parole, the accuracy is 97.1, whereas it is 93.7 with no threshold. It is not surprising that a method based on distributional features suffers when the absolute frequencies approach 1.

[9]http://korpus.dsl.dk/korpus2000/.

[10]http://www.id.cbs.dk/ mtk/treebank/

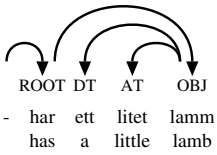

Figure 1: Talbanken05 annotation

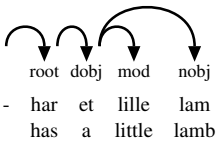

Figure 2: DDT annotation

	Animate			Inanimate		
	Prec	Rec	Fscore	Prec	Rec	Fscore
Swedish >10	81.9	64.0	71.8	96.4	98.6	97.5
Danish >10	74.5	45.5	56.5	95.5	98.7	97.1

Table 2: Precision, recall and F-scores for the two classes in the Swedish experiments, as well as the Danish experiments, evaluated through translational equivalents on the Talbanken05 data set.

yses of the two treebanks, hence parsers, are not completely isomorphic.

One point of difference between the two treebanks is in the head status of so-called functional categories, such as determiners and prepositions. Talbanken05 treats the nouns as heads with functional dependents, as illustrated in figure 1 where the determiner *ett* 'a' is a dependent of the noun *lamm* 'lamb'. The syntactic annotation in DDT, on the other hand, treats functional categories as heads with nominal dependents (nobj), as illustrated by figure 2, where the noun is a dependent of the determiner *et* 'a'. In extracting the distributional data for Danish, we wish to distinguish between various types of nominal argument relations such as subject, object and prepositional object. We therefore assign to the nouns the dependency relation of their head, e.g. the noun *lam* 'lamb' in figure 2 is assigned the `dobj`-relation of its determiner head.

With a few adjustments, we may thus employ the feature sets described in section 2 above to represent the Danish distributional data. With a frequency treshold of 10, to ensure sufficient distributional data, we end up with 18240 noun lemmas for classification. We apply the Swedish classifier to the Danish distributional data, resulting in a total of 16692 inanimate instances (91.5%) and 1548 animate instances (8.5%).

4 Evaluation

Evaluation of the resulting classification is not entirely straightforward, due to the fact that we do not have a Danish gold standard. Whereas this fact formed part of the motivation for this work, it also poses a challenge when we wish to evaluate the resulting classifier.

4.1 Evaluation through translation

If we assume that central semantic properties, such as animacy, do not differ between translational equivalents, we may use the Swedish gold standard annotation in order to evaluate the Danish classification.

We compile a Danish-Swedish lexicon from freely available, on-line resources.[11] The resulting dictionary contains a total of 5885 Danish-Swedish word pairs.[12] With this resource, we find a Swedish translation for 2555 of our classified Danish noun lemmas (18240 in total). Out of the set of classified Danish lemmas with a Swedish translation, 978 noun lemmas furthermore have a gold standard animacy annotation in the Talbanken05 data set. In the resulting gold-standard, the proportion of inanimate instances is 92.1, giving us a baseline for evaluation.[13]

The method for evaluation clearly only gives us an evaluation for a small subset of our classified lemmas. Even so, it might still give us a reasonable idea about the general quality of the ported classifier.

4.1.1 Results

The accuracy of the classifier when evaluated against the translated Talbanken05 data is 94.5, which constitutes a 30.3% reduction in error rate compared to the baseline. We find that the acquired classification furthermore is significantly better than the baseline ($p<.001$).[14]

The result for Danish is similar to the result obtained for the Swedish nouns of same frequency (95.1). Recall however that the Swedish and Dan-

[11] The free dictionaries project at http://www.dicts.info/ and dictionaries found at http://www.danska-svenska.se and http://dictionary.japplis.com/danish-swedish.html

[12] The lexicon consists of all types of word classes, not only nouns. Furthermore, the on-line resources from which the lexicon was compiled have largely been constructed automatically, hence are by no means perfect.

[13] Note however, that the baseline does not necessarily reflect the true distribution of animate vs. inanimate instances in Danish. The fact that it is higher than in the Swedish data is an indication that it might be artificially high due to arbitrary properties of the dictionaries used for evaluation.

[14] For calculation of the statistical significance of differences in the performance of classifiers tested on the same data set, McNemar's test is employed.

ish classifiers are evaluated on different data sets. The Swedish classifier is evaluated on the total Talbanken05 data set presented in table 1,[15] whereas the Danish classifier is evaluated on the nouns in this data set for which there is a Danish-Swedish translation *and* more than 10 instances in the Korpus2000 corpus.

With respect to the classes of 'animate' and 'inanimate', table 2 reports the class-based measures of precision, recall and F-score for the Swedish and Danish classifiers. The baseline F-score for the animate class is 0, and a main goal in classification is therefore to improve on the rate of true positives for animate instances, while limiting the trade-off in terms of performance for the majority class of inanimates, which start out with F-scores approaching 100. We find that the performance for the minority class of 'animate' is generally lower than in the Swedish results. Like the Swedish results, however, we find that the classifier is consevative in terms of assignment of the minority class of 'animate' and shows a fairly high precision (74.5), combined with a lower recall (45.5) for this class. For the majority class of 'inanimate', the performance for the two languages are highly similar, with F-scores between 97.1-97.5.

5 Conclusions and future work

The porting of a classifier for the semantic property of animacy, trained on distributional frequencies for noun lemmas, turns out to work quite well for the highly related language-pair Swedish-Danish. Distributional data describing the general morphosyntactic distribution of nouns was extracted for the new language, Danish, and a classifier trained on corresponding data for the source language, Swedish, was then applied. We evaluated the resulting classification by means of translation to the gold standard annotation in Swedish and found that the resulting classifier gave significant improvements over a majority baseline. We obtain an accuracy of 94.5 on the Danish evaluation data set, constituting a 30.3% reduction of error rate. Using only a large, automatically annotated corpus for the second language, Danish, we were able to obtain animacy annotation for a total of 18240 noun lemmas. This clearly gives us a better coverage than one what one might expect from an approach relying on translation by means of lexical resources.

In terms of future work, we are interested in the application of a similar methodology (i) to other language pairs, both highly related, e.g., German-Dutch, Spanish-Italian, and less related ones, and (ii) to other semantic classification tasks, such as verb or adjective classification.

[15] With the restriction that it occurs more than 10 times in the Parole corpus.

References

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Peter de Swart, Monique Lamers, and Sander Lestrade. 2008. Animacy, argument structure and argument encoding: Introduction to the special issue on animacy. *Lingua*, 118(2):131–140.

Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. 2005. A maximum entropy model of subject/object learning. In *Proceedings of the 2nd Workshop on Psychocomputational Models of Human Language Acquisition*.

Johan Hall. 2003. A probabilistic part-of-speech tagger with suffix probabilities. Master's thesis, Växjö University, Sweden.

Richard Johansson and Pierre Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. In *Proceedings of COLING/ACL*.

Matthias Trautner Kromann. 2003. The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT)*.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING)*, volume 2, pages 768–774.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Procs. of the 40th Meeting of the Association for Computational Linguistics (ACL'02)*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006a. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006b. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.

Constantin Orăsan and Richard Evans. 2007. NP animacy resolution for anaphora resolution. *Journal of Artificial Intelligence Research*, 29:79–103.

Lilja Øvrelid and Joakim Nivre. 2007. When word order and part-of-speech tags are not enough – Swedish dependency parsing with rich linguistic features. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*.

Sebastian Pado and Mirella Lapata. 2005. Cross-lingual projection of role-semantic information. In *Proceedings of HLT/EMNLP 2005*.

Suzanne Stevenson and Eric Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 71–78.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206