Evaluating automatic cross-domain Dutch semantic role annotation

Orphée De Clercq[1,2], **Veronique Hoste**[1,3], **Paola Monachesi**[4]

LT3, Language and Translation Technology Team, University College Ghent[1]
Groot-Brittanniëlaan 45, 9000 Ghent Belgium
Department of Applied Mathematics and Computer Science, Ghent University[2]
Krijgslaan 281 (S9), 9000 Ghent Belgium
Department of Linguistics, Ghent University[3]
Blandijnberg 2, 9000 Ghent Belgium
Uil-OTS, Utrecht Institute of Linguistics OTS, Utrecht University[4]
Trans 10, 3512 JK Utrecht The Netherlands
orphee.declercq@hogent.be, veronique.hoste@hogent.be, P.Monachesi@uu.nl

Abstract

In this paper we present the first corpus where one million Dutch words from a variety of text genres have been annotated with semantic roles. 500K have been completely manually verified and used as training material to automatically label another 500K. All data has been annotated following an adapted version of the PropBank guidelines. The corpus's rich text type diversity and the availability of manually verified syntactic dependency structures allowed us to experiment with an existing semantic role labeler for Dutch. In order to test the system's portability across various domains, we experimented with training on individual domains and compared this with training on multiple domains by adding more data. Our results show that training on large data sets is necessary but that including genre-specific training material is also crucial to optimize classification. We observed that a small amount of in-domain training data is already sufficient to improve our semantic role labeler.

Keywords: corpus annotation, semantic role labeling, cross-domain

1. Introduction

In the last decade a lot of time and effort has been devoted to creating new resources for Dutch initiated by the STEVIN programme[1]. These efforts have been united in one last project: the SoNaR corpus, which comprises 500 million words of written Dutch text including various text genres (Reynaert et al., 2010). In addition, a core subset of one million words[2] is enriched with four semantic layers: named entities, coreference relations, semantic roles and spatio-temporal relations (Schuurman et al., 2010).

In this paper we discuss the annotation process of 500K and we report on our approach to automatically annotate Dutch semantic roles based on these manually verified data.

For semantic role labeling (SRL) – the task of automatically defining *who did what to whom when* – the link between syntax and semantics is crucial and has been underlined many times (Punyakanok et al., 2008). A first choice is to decide which basic syntactic representation to follow. For Dutch, Monachesi et al. (2007) were among the first to choose dependency over constituent syntax because of its rich syntactic information and ability to provide very useful information on the relation between parts of a sentence such as grammatical functions.[3] After both the CoNLL 2008 and 2009 tasks were devoted to this subject, dependency structures now seem common practice.

Based on these findings it was decided to use dependency syntax as starting point for the annotation of the semantic roles in the SoNaR subcorpus, a main advantage being also that we were able to rely on manually verified Dutch dependency structures. These were annotated as part of the LassyKlein Corpus project[4].

Another advantage is the corpus's rich text type diversity in that it comprises six distinct genres: administrative texts, autocues, texts treating external communication, instructive texts, journalistic texts and wikipedia. This enabled us to experiment with the portability of an existing labeler for Dutch (Stevens et al., 2007). Instead of focusing on the labeler and features itself, we closely investigate the effect of training on a more diverse data set to see whether merely varying the genre or amount of training data optimizes performance. Similar work was done by De Clercq et al. (2011), focussing on Dutch coreference resolution.

In this paper, we report on a set of experiments in which both in-domain and out-of-domain data are used training material for the semantic role labeler, we show that training on large data sets is necessary but that including genre-specific information is also crucial to optimize classification. Given the performance differences between the different genres we evaluated on, we also performed an error analysis of the lexical sparseness of the predicates within each genre, which revealed that a low number of unique predicates results in better performance for that particular genre. The genre's generalization power, however, does not benefit from this weak representation when applied to a diverse data set.

The paper is structured as follows. In Section 2 we discuss the methodology followed to annotate and automati-

[1] http://taalunieversum.org/taal/technologie/stevin/english/

[2] This subcorpus, SoNaR1, will be distributed by the Dutch HLT-agency as an integral part of SoNaR.

[3] For a full discussion we refer to Johansson and Nugues (2008)

[4] Available at: http://www.inl.nl/tst-centrale/nl/

cally label semantic roles based on manually verified data. The corpus, experimental set-up and cross-domain experiments are presented in Section 3. We examine the results and perform an error analysis in Section 4, to end with some conclusions and prospects for future work (Section 5).

2. Towards automatic semantic role annotation

SoNaR presents the first corpus where one million Dutch words are annotated with semantic roles: 500K has been completely manually verified and used as training material to automatically label the remaining 500K. The actual semantic roles are added as an additional layer on top of manually verified dependency trees.

Based on the positive findings in Monachesi et al., (2007) an existing labeler (Stevens et al., 2007) was retrained on a small set of roughly 2,000 manually verified sentences to pretag semantic roles as a starting point. After this, experiments were conducted to optimize performance and each time retrain on a more substantial data set.

The semantic roles of 500K have been manually verified following the guidelines developed by Trapman and Monachesi (2006) who adapted the PropBank guidelines (Babko-Malaya, 2005) so as to handle Dutch text. Only framefiles of predicates (verbs) were labeled and instead of creating new Dutch framefiles, everything was mapped onto English PropBank frames. As annotation environment TrEd[5] was used.

In order to validate our approach, two qualitative error analyses were performed: one after an initial 50K had been manually verified and the second after 300K had been checked. The first analysis revealed that for our labeler especially higher numbered arguments (Arg3, Arg4) are more difficult to label (Example 1). Whereas for the annotators, deciding which PropBank framefile applied to which Dutch verb sometimes proved problematic (Example 2), mainly because multiple translations to English are often possible regardless of the context.

Example 1. De agrarische productie — stijgt — van 20 [Arg3] — naar 25% [Arg4]. (*Engl. The agrarian production — has increased — from 20 [Arg3] — to 25% [Arg4].*

Example 2. Steun voor het onderzoek dat Pronk moet uitvoeren. (*Engl. Aid for the research Pronk has to conduct, execute, perform, carry out, ...*)

Based on these findings, we decided to also include the English PropBank framefile as an additional attribute in our data. Here is an example of a Dutch sentence annotated with semantic roles and English PropBank frame:

Example 3. Ik [Arg0] — heb — de hoop [Arg1] — nooit [ArgM-NEG] — verloren [PRED, pbframe: lose.02]. (*Engl: I never lost hope.*)

In order to ensure consistency among annotators, they were all provided with a joint list in which each annotator indicated which English framefile he/she used for which Dutch verb and all annotators got the instruction to always consult this list prior to annotation. If they did not know or were unsure about a certain translation they could choose a dummy label.

This allowed us to verify whether transferring English PropBank frames to Dutch verbs was a valid approach. We investigated this by counting how many times the annotators chose this dummy label after 300,000 words had been annotated. This set contained 21,419 predicates and we saw that for 503 predicates a dummy label was chosen (about 2%). After examining these cases we see that most of these verbs can be reconciled with a particular PropBank framefile, thus leaving us with less than one percent of predicates for which no PropBank frame exists. These are mostly idiomatic expressions for which no English counterpart exists (Example 4).

Example 4. Alexandre Thelahire werd 48 uur gegijzeld. (*Engl. Alexandre Thelahire was taken hostage for 48 hours*).

These findings provide evidence for PropBank's cross-lingual validity which was already analyzed in close detail for French by Van der Plas et al. (2010).

3. More data versus genre-specific data

Ever since the seminal work of Gildea and Jurafsky (2002), semantic role labeling is perceived as a task in which two steps are performed: argument identification and argument classification. Previous research has shown that for the first step syntactic knowledge is important whereas the second one necessitates more semantic information (Pradhan et al., 2008).

Since we have golden dependency structures to start with for this basic step of argument identification, we were able to focus more on argument classification for our experiments. Because of the unique composition of the SoNaR subcorpus which is distributed over six distinct text genres, we were able to experiment with these multiple genres and actually examine the difference in performance when trained on these various domains.

This issue of cross- or open-domain semantic role labeling has already been explored for English (Johansson and Nugues, 2008; Pradhan et al., 2008) focussing on training an existing labeler on one domain and testing it on a different domain. Pradhan et al.'s (2008) main finding is the poor generalization power of lexical features which inspired further work by, for example, applying semi-supervised learning to increase lexical expressiveness (Croce et al., 2010). The focus in this paper differs in that we take it one step further by not only training on one genre and testing it on another genre but the corpus's rich diversity gave us the possibility to train on different individual genres and compare this with a combination of these genres. We aim to find out whether training on a more diverse data set comprising different domains results in a more robust labeler.

3.1. Data sets

The completely manually verified data set consists of 500,850 tokens and can be divided into six distinct text genres based on its origins. In the administrative genre (ADM)

[5] http://ufal.mff.cuni.cz/ pajas/tred/

reports, speeches and minutes of meetings are included; the autocues genre (AUTO) consists of written newswire. Another genre, referred to as external communication (EXT), represents website material, press releases and newsletters. The instructive texts (INST) genre includes manuals, patient information leaflets and procedure descriptions whereas the journalistic genre (JOUR) consists mainly of newspaper articles. Finally, the sixth genre has data originating from Dutch wikipedia (WIKI).

In Table 1, some corpus statistics are presented. Besides the number of tokens, sentences and predicates we also mention the average number of predicates per sentence because these might hint at a more complex sentence structure.

We immediately see that the genres are quite unbalanced. That is why we decided to also experiment with balanced data sets, each amounting to 50K. This is discussed in closer detail in the next section.

Data sets	#Tokens	#Sentences	#PRED	Avg./Sent.
ADM	63,063	3,422	4,192	1.22
AUTO	94,371	6,438	7,780	1.21
EXT	98,618	5,785	7,185	1.24
INST	60,959	3,069	4,012	1.31
JOUR	89,420	4,657	7,303	1.57
WIKI	94,419	6,047	6,486	1.07
TOTAL	500,850	29,418	33,256	1.13

Table 1: Data statistics indicating the number of tokens, sentences and labeled predicates present in each text genre of the manually verified SoNaR 1 subcorpus plus the average number of predicates per genre

3.2. Experimental set-up

For all experiments we used a semantic role labeler that was originally developed by Stevens et al. (2007) and which was further improved during the SoNaR project. The system follows a pair-wise approach in which each instance contains features of a predicate and its candidate argument. To prevent overfitting, only siblings of the verb in the dependency structure are considered candidate arguments.

A number of features are extracted from the predicate-argument pairs to describe their relation. As features, our system uses a standard feature set for dependency-based semantic role labeling and we were able to include gold preprocessing information thanks to the available golden dependency trees.

Three properties of the predicate are described:

- the verb's lemma,
- the part-of-speech tag,
- the voice (active or passive).

The candidate argument features encode:

- the c-label, i.e. the category label of the possible argument (whether it is a noun phrase, a prepositional phrase,...),
- the dependency or d-label (is the argument a subject, modifier,...),
- a binary feature indicating whether the argument is positioned before or after the predicate,
- the head(lemma) together with its corresponding part-of-speech tag,
- if an argument consists of multiple words, the first and last word together with their part-of-speech tags are also included,
- the CAT/POS pattern, i.e. the left-to-right chain of d-labels of the candidate argument and its siblings,
- the REL pattern, i.e. the left-to-right chain of c-labels of the candidate argument and its siblings,
- the CAT+REL pattern which is the c-label of the argument concatenated with its d-label.

For all experiments we used Timbl version 6.3 (Daelemans et al., 2010) with default parameter settings. Results are evaluated by calculating precision, recall and F-measure. We each time present the overall scores for argument classification.

Three sets of experiments were conducted:

1. In the first experiment, we evaluated the semantic role labeler on in-domain data using 10-fold cross validation. We conducted two sub-experiments, one in which the original data distribution (see Table 1) was kept and a second for which we selected a random sample of 50K per genre.
2. In the second set, we explicitly focused on the **cross-domain experiments**. In order to rule out data set size as a disturbing factor which could bias results, we continued working with the 50K data sets per genre. The main objective of the experiments was to find out what works best for our classifier: training on in-domain data or on a more diverse data set incorporating a variety of genres. In order to do so, we conducted a set of 10-fold cross-validation experiments on the 50K data sets. In order to allow for comparison, the 5K test set partitions were kept constant over all experiments. In a first experiment, the classifier was trained on the 45K in-domain training partitions and tested on the relevant 5K test partitions. In a second experiment, the robustness of the classifier was evaluated by exclusively training the classifier on out-of-domain data. In a final step, also in-domain data was included in the training data.
3. The third set of experiments adds more data to the labeler. In order to do so, we used the balanced data set, in which each genre is equally represented. In a first experiment, we included out-of-domain data (5 genres * 50K = 250K) and tested the performance of the classifier on the 5K test partitions mentioned in the previous set of experiments. In a final step, also in-domain data was included in the training data (5 genres * 50 K + 1 genre * 45K = 295K).

4. Results and Discussion

The results of the first set of experiments are presented in Table 2. When we compare the in-domain results of training on all data versus a balanced subset of 50K, we can observe that there only exists a (modest) difference in performance when a substantial amount of training data is added, as in the case of the AUTO, EXT, JOUR and WIKI data sets. The smaller data sets, viz. ADM and INST, on the other hand, seem to benefit from a reduction in training material. A more qualitative analysis should reveal why this is the case. On both data sets, the best results are obtained for the instructive genre.

In the second set of experiments we ruled out data set size as a potentially biasing factor and focused on the 50K data sets. The results of these experiments are presented in Table 3. For ease of comparison, we listed the experimental results on the balanced in-domain data sets (50K per genre) as listed in the second part of Table 2, as first columns in Table 3.

Whereas in the first set of experiments, we contrasted how the labeler would perform in case it was trained on a data set which was specifically tailored to the test set at hand, the second experiment aimed at completely the opposite. In the second experiment, for which the results are given in columns 5 to 7 in Table 3, we investigated how the labeler which was trained on data from other genres than the genre to be tested on, would perform on this held-out genre for which it did not receive any training evidence. This would allow us to draw some conclusions on the robustness of the semantic role labeler. We clearly observe a drop in performance on all genres, ranging between 2 and 4%. Although adding no in-domain information at all for training the labeler seems to harm its performance, the performance drops remain modest (except maybe for the INST genre). This led us to additional experiments in which we also included a small sample, viz. 1/6 of the 50K, of in-domain training data instead of only focussing on other genres. The results for these In- & out-of-domain experiments (columns 8 to 10) show an improvement for all genres, also outperforming the in-domain experiments. This might mean that in order for our labeler to perform better on a particular genre a small amount of in-domain training data already helps.

These findings had to be corroborated on larger data sets though, and that is why we decided to perform a third round of experiments where more data is added to our labeler. Similar experiments were performed but we now decided to include all available data from the other genres. We again made a subdivision between working only on out-of-domain data and including both in- and out-of-domain data, the same test sets were used as in the second round. The results of these final experiments are presented in the lower part of Table 3. We see that when including 250K of out-of-domain training data only three genres seem to benefit from this (EXT, JOUR and WIKI). If we also include a small amount of in-domain information, however, we see that for all genres the best results are achieved. This further strengthens our assumption that including genre-specific training information is necessary and we clearly see that a small amount can already account for better results.

4.1. Error Analysis

Overall, in the final round of experiments we observed that the highest F-score is reached for the instructive text genre (79.18), whereas the other genres have all more or less the same performance. In the other experiments, the instructive genre was also always the most performant one. This necessitated a closer look at this specific genre. If we look at its data statistics in Table 1, we see nothing extraordinary: it contains about 1.3 predicates per sentence which is not particularly high nor low. Looking at the origins of the instructive text genre, however, we see that it contains manuals, patient information leaflets and procedure descriptions which might mean that the text material itself is not very diverse. This was further investigated by looking at the lexical predicate sparseness within each genre, i.e. how many predicate types are represented within one genre. The results are presented in Figure 1[6].

Overall, the highest number of distinct predicates can be found in the journalistic genre (1,109) and the lowest number within the instructive genre (475). Due to the low variability in the predicates, our ten-fold cross validation set-up might have benefited from the fact that many predicates from the test data were also represented in the training data and thus less difficult to classify. This finding, however, does not prevent that the best results were achieved while also including a large amount of out-of-domain training data.

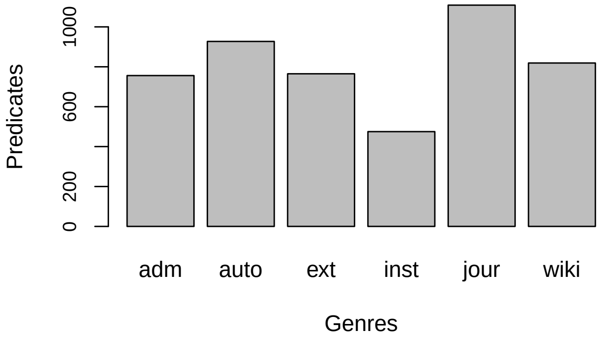

Figure 1: the number of unique predicates within each genre

In order to further investigate this we did two additional cross-domain experiments by training on the instructive genre (50K) and testing on the journalistic genre (50K) and vice versa. We see that the instructive genre's generalization power is weaker than that of the journalistic texts (F-measure of 66.70 versus one of 70.88). This further underlines the necessity of including enough out-of-domain data together with a small amount of in-domain data when one wishes to have a robust classifier.

[6] These data have been derived from the balanced data sets (50K) but the same tendency was perceived in the entire corpus

	In-domain all (500K)			In-domain balanced (300K)		
Data sets	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
ADM	72.40	71.35	71.70	73.21	72.63	**72.92**
AUTO	72.75	72.40	**72.57**	72.63	72.19	72.40
EXT	72.45	71.80	**72.12**	72.12	71.11	71.61
INST	76.12	75.72	75.92	77.53	77.12	**77.32**
JOUR	73.51	72.93	**73.22**	72.14	71.36	71.75
WIKI	74.22	73.27	**73.74**	73.05	71.85	72.44

Table 2: Results of training on in-domain data using 10-fold cross validation

	In-domain (45K vs 5K)			Out-of-domain (45K vs 5K)			In- & out-of-domain (45K vs 5K)			Out-of-domain (250K vs 5K)			In- & out-of-domain (295K vs 5K)		
Data sets	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
ADM	73.21	72.63	72.92	70.26	70.00	70.13	74.05	73.90	73.97	72.48	72.75	72.61	74.67	74.69	**74.68**
AUTO	72.63	72.19	72.40	70.16	68.28	69.21	73.10	72.19	72.64	73.14	72.00	72.57	74.65	73.73	**74.19**
EXT	72.12	71.11	71.61	70.92	70.18	70.55	73.31	72.48	72.89	73.57	73.01	73.29	74.48	73.71	**74.09**
INST	77.53	77.12	77.32	73.54	73.09	73.31	77.88	77.64	77.76	75.61	76.06	75.83	79.18	79.18	**79.18**
JOUR	72.14	71.36	71.75	71.52	70.65	71.08	72.98	72.35	72.66	73.47	72.92	73.19	74.47	73.89	**74.18**
WIKI	73.05	71.85	72.44	71.02	69.60	70.30	73.65	72.35	72.99	74.52	73.71	74.11	75.87	75.03	**75.45**

Table 3: Results of training on in-domain data and while adding more data (both in- and out-of-domain)

5. Conclusion and Future Work

In this paper we have presented the work carried out to annotate one million words of Dutch text with semantic roles following the PropBank scheme. A dependency-based semantic role labeler was trained to speed up the manual annotation task and PropBank's cross-lingual validity was underlined for Dutch. Our main objective was to automatically label 500K with high precision.

The corpus's rich text genre diversity allowed us to experiment with an existing labeler by training it on each genre individually and on a substantial data set comprising all these different genres. The experiments reveal that training on large data sets is necessary but that including genre-specific training material is also crucial to optimize classification. Moreover, we observed that a small amount of in-domain training is already sufficient. An error analysis of the lexical sparseness of the predicates within each genre revealed that a low number of unique predicates results in better performance for that particular genre. A particular genre's generalization power, however, does not benefit from this weak representation. This further underlines the importance of also including enough out-of-domain data when the main objective is to have a robust labeler.

Future work includes adding new lexical and semantic features to our labeler, most notably features available from the other SoNaR layers to verify the upper bound of our system. In a final step we would also like to investigate the performance of our labeler when using automatically parsed syntactic dependency structures instead of manually verified ones.

Acknowledgements

The work presented in this paper was made possible by the STEVIN programme of the Dutch Language Union within the framework of the SoNaR project under grant number STE07014 and the Flemish EWI Council in the framework of the Stylene project.

6. References

O. Babko-Malaya. 2005. Propbank annotation guidelines. Technical report.

D. Croce, C. Giannone, P. Annesi, and R. Basili. 2010. Towards open-domain semantic role labeling. In *Proceedings of the ACL 2010 Conference*, Uppsala, Sweden.

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2010. Timbl: Tilburg memory based learner, version 6.3, reference guide. ILK Research Group Technical Report Series 10-01, Tilbug University.

O. De Clercq, I. Hendrickx, and V. Hoste. 2011. Cross-domain dutch coreference resolution. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

D Gildea and D. Jurasky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288.

P. Johansson and P. Nugues. 2008. The effect of syntactic representation on semantic role labeling. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, UK.

P. Monachesi, G. Stevens, and J. Trapman. 2007. Adding semantic role annotation to a corpus of written dutch. In *Proceedings of the Linguistic Annotation Workshop*, Prague, Czech Republic. ACL.

S Pradhan, W. Ward, and J. Martin. 2008. Towards robust semantic role labeling. *Computational Linguistics*, 34:289–310.

V Punyakanok, D. Roth, and Y. Wen-tau. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34:257–287.

M. Reynaert, N. Oostdijk, O. De Clercq, H. van den Heuvel, and F. de Jong. 2010. Balancing sonar: Ipr versus processing issues in a 500-million-word written dutch reference corpus. In *Proceedings of LREC'10*, Valletta, Malta. ELRA.

I. Schuurman, V. Hoste, and P. Monachesi. 2010. Interacting semantic layers of annotation in sonar, a reference corpus of contemporary written dutch. In *Proceedings of LREC'10*, Valletta, Malta. ELRA.

G. Stevens, P Monachesi, and A. van den Bosch. 2007. A pilot study for automatic semantic role labeling in a dutch corpus. In *Selected papers from the seventeenth CLIN meeting*, Utrecht, The Netherlands. LOT Occasional Series 7.

J. Trapman and P. Monachesi. 2006. Manual for semantic annotation in d-coi. Technical report, Utrecht University, Uil-OTS.

L. Van der Plas, T. Samardžić, and P. Merlo. 2010. Cross-lingual validity of propbank in the manual annotation of french. In *Proceedings of the ACL 2010 Conference*, Uppsala, Sweden.