Supertagging for Domain Adaptation: An Approach with Law Texts

Kyoko Sugisaki*
Institute of Computational Linguistics, University of Zurich
Andreasstrasse 15, 8050 Zurich, Switzerland
sugisaki@cl.uzh.ch

ABSTRACT

In this paper, we present a German supertagger that analyses syntactic functions in linear order. We apply a statistical sequential model, conditional random fields (CRF), to Swiss law texts, in a real world scenario in which the training data of the domain is missing. We show that the small amount of in-domain training data that was informed by linguistic hard and soft constraints and domain constraints achieved a label accuracy of 90% in the domain data, thus outperforming state-of-the-art parsers.

CCS Concepts

•**Computing methodologies** → Language resources;

Keywords

Natural language processing; annotation of law texts

1. INTRODUCTION

Recently, interest has increased in using natural language processing (NLP) to apply statistical methods to domain texts, particularly statistical parsing [1, 2, 3]. Statistical parsers learn the language patterns of the training data (mainly newspaper texts). However, the learned models are usually not general enough to be applied to other domains and text types [3, 4]. Under these circumstances, the best approach is the manual annotation of a large amount of new domain data in which a parser can be trained. However, this is also the most cost-intensive solution. The second approach is to apply semi-supervised domain adaptation methods, such as self-training [1, 2, 5, 6, 7, 8] and co-training [5, 9, 10]. These approaches are used to improve the performance of statistical parsers without the need to annotate in-domain data manually. These semi-supervised methods have the advantage that they can be used in other domains. However, the results of previous research on using these approaches are mixed. Both positive [1, 2, 7, 8, 10] and negative results [6, 9] have been reported.

In this work, we develop a method for predicting dependency grammar functions with high accuracy under the constraint that only minimal in-domain manual data annotation can be undertaken. Our domain texts are Swiss law texts written in German. We advance the state-of-the-art supervised dependency grammar parsing by providing the linear classifier conditional random fields (CRF) [11] with multivariate linguistic information that allows for the handling of complex syntax in the linear context. In particular, we investigate a method to incorporate two types of linguistic information in a statistical model: hard and soft grammar constraints. Hard constraints are rules that cannot be disregarded without violating grammaticality. In contrast, the violation of soft-constraint rules does not lead to ungrammaticality. Hard constraints, for example, are morphological rules, such as 'nouns agree with determiners in number, gender, and case" or argument structure rules, such as 'there is only one subject in a clause." Soft constraints are probabilistic grammar rules. An example is 'animate entities are more likely to be subject".

The paper is organized as follows. We start by describing the design of our CRF-based supertagger. Then, we describe our experiments to train the sequential model for the domain of Swiss German-language law texts. We conclude by evaluating the results of our experiments with the CRF-based supertagger.

2. CRF-BASED SUPERTAGGING

We develop a CRF-based German supertagger to assign 17 dependency grammar labels to nouns, pronouns and prepositions in a clause (cf. tb. 1). In this section, we briefly introduce the Conditional Random Fields (CRF) and define the notion of a sequence and a set of features for the task.

2.1 CRF

In this subsection, we outline the algorithm of CRF [11, 16]. CRF is a random field for conditional probability $P(y_{1:n}|x_{1:n})$, where $x_{1:n}$ is an input sequence $x_1 \dots x_n$ and $y_{1:n}$ is an output sequence $y_1 \dots y_n$. To calculate the conditional probability, CRF makes use of the maximum entropy model and normalizes the probability globally in a sequence:

$$P(y_{1:n}|x_{1:n}) = \frac{1}{Z(x_{1:n})} \exp\left(\sum_{n=1}^{N}\sum_{d=1}^{D} w_d f_d(x_{1:n}, y_n, y_{n-1}, n)\right)$$

The term $Z(x_{1:n})$ sums over all possible values of the se-

*The project is funded under SNSF grant 134701.

DOI: 10.1145/1235

SUBJ(Subject)	OBJA (Direct object)	OBJD (Indirect object)	OBJG (Genitive object)
GMOD (Genitive modification)	APP (Apposition)	PN (Complement of preposition)	ROOT (Root)
KON/CJ (Complement of conjunction)	GRAD (Grade)	ZEIT (Time)	EXPL(Expletive)
PRED (Predicate)	PAR (Parenthesis)	PP (Adjunct preposition)	OBJP (Prepositional object)

Table 1: Target dependency labels

#	Linguistic cue	Context window	Description
1	Word form	[-2] [-1] [0] [+1] [+2]	Word form
2	Word character	[0]	Capitalization, numeric and alpha-numeric characters and 2 - 4 suffix characters of full nouns
3	Child string	[0]	(1) Article or the ending of strongly declined adjectives (e.g. *-e, -em, -er*) or (2) the word forms of complements of prepositions
4	Child type	[0]	(1) The types of dependents of full nouns: Definiteness or adjective, numeral or bare (2) The types of dependents of prepositions: full nouns, pronouns or relative pronouns
5	Coarse POS	[-1] [0] [+1]	Full nouns, pronouns, relative pronoun s or prepositions
6	Topological fields	[-1] [0] [+1]	Vorfeld, mittelfeld, nachfeld or without field [12].
7	Animacy	[0]	Person, organization or animal: GermaNet [13], Gertwol [14] and a list of professions (37,494 entities).
8	Brown clustering	[-1] [0] [+1]	Top 4 and full Brown word clusters (2 features) [15]
9	Contexts	[-1] [+1]	Coordinating and comparative conjunctions, punctuations, quotation mark, left and right brackets and adverbs
10	Predicate	[0]	Predicates
11	Voice	[0]	Passive or active voice

Table 2: Baseline features(context: [0] = current token, [-1] = previous token, [+1] = next token)

quence $y_{1:n}$:

$$Z(x_{1:n}) = \sum_{y_{1:n}} \exp\left(\sum_{n=1}^{N}\sum_{d=1}^{D} w_d f_d(x_{1:n}, y_n, y_{n-1}, n)\right)$$

2.2 Sequence

A CRF model learns the parameters and decodes the output based on a given sequence of input tokens. To model the linearisation of arguments including grammatical functions in German, we define a sequence in two ways. First, a sequence contains only nouns, pronouns and prepositions. In other words, we skip the tokens in which we are not interested and create a sequence containing only tokens relevant for the prediction of grammatical functions. Second, a sequence corresponds to a clause, instead of a sentence. A clause is a linguistic unit in which the arguments of a verb interact uniquely with each other function.

2.3 Features

Features are key probabilistic (or soft) linguistic indicators that may be useful to predict of dependency grammar labels. Table 2 provides an overview of the 11 features, which are morphological, syntactic, semantic, and pragmatic linguistic cues for the linearisation of grammatical functions and head-modifier dependency grammar relations (e.g. apposition).[1]

3. EXPERIMENTS

In this section, we describe our experiments to train the sequential model for the domain of Swiss law texts. Our experimental setting is a real world scenario, where the syntactically annotated high-quality data of own domain data does not exists, but a large amount of that of out-of-domain data is available.

The data sets used for training in our experiments are (1) a large amount of data from TüBa D/Z (Henceforth: TüBa), a German newspaper corpus, as out-of-domain data and (2) a small amount of data (300 sentences) from Swiss law texts, as in-domain data. As our test domain data, we used another 200 sentences from Swiss law texts. These data sets are randomly selected from the collection of Swiss law texts. Then, we have annotated them manually for the experiments. The features presented in the previous section are used as base features in the experiments.

[1]We used the Brown clustering (#8 in tb. 2) implemented by P. Liang: https://github.com/percyliang/brown-cluster

3.1 Adapting Out-of-Domain Training Data

In this subsection, we explore how the existing large amount of annotated out-of-domain training data (here: TüBa) could be adapted to the legislative domain.

In the experiment, we first use a large amount of data from TüBa for training. Thus, we train a CRF model on 60% of TüBa (TüBa60Train, 700,888 tokens) as baseline model. We then select this out-of-domain training data for the domain adaptation. We use two different types of data selection methods: (1) Cross-entropy based data selection (2) Linguistically motivated data selection.

In cross-entropy-based data selection, sentences that are similar to the target legislative domain are automatically extracted from a large amount of out-of-domain data. Cross entropy is a variant of perplexity that is used to compare different probability models. It is measured as follows [17]:

$$H(w_1 \ldots w_n) = -\frac{1}{N} logP(w_1 \ldots w_n) \quad (1)$$

Cross entropy is approximated by computing the log probability of a sequence $w_1 \ldots w_n$, normalized by the length of the sequence. In this experiments, we trained word-based and POS-based 4-gram language models with Kneser-Ney back-off on the TüBa training data (39,313 sentences, 700,888 tokens) and the same amount of law text data (25,901 sentences, 700,974 tokens). We used the language model, Berkeley LM [18].

To measure the similarity of out-of-domain data to the targeted in-domain legislative data, We used the following two measurements: (1) Ranking-based entropy score (2) Difference-based entropy score.

Ranking-based entropy scoring is a measurement of how

Training: method, % of TüBa60Train	Test:Law200Test
None 100%	**84.56**% (82.12%)
Word-Entropy-Ranking 25%	84.17% (75.47%)
Word-Entropy-Difference 25%	84.85*% (63.20%)
POS-Entropy-Ranking 25%	83.83% (82.43*%)
POS-Entropy-Difference 25%	84.89*% (**83.40***%)

Table 3: Out-of-domain data selection: label accuracy all (grammatical functions)

#	Training	Test	
A	TüBa60Train	84.56%	(82.12%)
B	TüBa60NoLawNoise (TBNLN)	**85.81**%	(**83.73**%)
C	TBNLN + Law100Train	86.00%	(82.73%)
D	TBNLN + Law200Train	86.15%	(83.03%)
E	TBNLN + Law300Train	**86.53**%	(83.47%)
F	TBNLN + Law300Train with case	88.61%	(87.96%)
G	TBNLN + Law300Train with dep	92.57%	(**90.75**%)
H	TBNLN + Law300Train with case/dep	**92.86**%	(90.38%)

Table 4: CRF models with soft and hard constraints: Label accuracy all (grammatical functions)

surprisingly out-of-domain sentences are encountered based on the experiences collected in the target-domain data, that is, the language model trained on target-domain data. The cross entropy scores were ranked by ordering from low to high. Out-of-domain TüBa sentences were assumed similar to the targeted legislative domain data if they were low in cross entropy, that is, less surprising. This method is compatible with [19] in which perplexity was used instead of cross entropy.

Difference-based entropy scores are a measurement of differences in entropy scores for out-of-domain sentences returned by a language model trained on target-domain data and by a language model trained on out-of-domain data. Out-of-domain sentences were considered similar to target-domain sentences if the difference of entropy scores between those two models was small. The method is based on [20].

In the experiment, we trained CRF models on 25, 50, 75, 85, and 95% of the TüBa training set (TüBa60Train). We tested the models on the test set of the law texts. A striking result was that 25% of the training data were competitive with 100% of the training data with regard to overall label accuracy and grammatical functions (cf. tb. 3).

In addition, we conducted an experiment with a TüBa training data set that was selected based on the following linguistic observations: The law texts did not contain verbless clauses, fragmental clauses, or parenthetical clauses. Therefore, we simply remove these types of clauses by getting rid of clauses that contained dependency labels `ROOT` or `PAR` assigned to nouns or prepositions. The selected training set contains 56,296 clauses, which is 89% of the whole TüBa training set (TüBa60Train). We train a CRF model on this training set (TüBa60NoLawNoise) and test the trained model on the test set of Swiss law texts (Law200Test). The model trained on TüBa60NoLawNoise outperformed the baseline model trained on TüBa60Train, i.e. the whole data set of the TüBa training data without the noise removal (#A in tb. 4), and the TüBa training data (25%) selected by the POS-based difference model (cf. tb. 3).

3.2 Adding In-Domain Training Data

Next, we add our in-domain training set into the domain-adapted training set, because the encoding of grammatical functions is also dependent on the type of texts. For example, in law texts, a provision is often subject (e.g. "Das Nähere bestimmt das Gesetz" 'the details, the act determines'). In order to integrate such domain information, we trained CRF models on TüBa60NoLawNoise combined with three in-domain data sets - 100 sentences ('Law100Train' in #C of tb. 4), 200 sentences ('Law200Train' in #D of tb. 4) and 300 sentences ('Law300Train' in #E of tb. 4).

Table 4 shows that only 300 sentences of annotated in-domain data improved the label accuracy of approximately 1% of the data in the previous experimental setting, that is, the use of only the out-of-domain training data.

	Rule-based		CRF(#E)		Hybrid(#H)	
	0.84	(1467, 275)	0.86	(1793,279)	**0.92**	(1924, 148)
SUBJ	0.82	(0.85, 0.81)	0.89	(0.91, 0.88)	**0.95**	(0.96, 0.95)
OBJA	0.68	(0.67, 0.68)	0.72	(0.69, 0.74)	**0.87**	(0.84, 0.90)
OBJD	0.34	(0.63, 0.24)	0.68	(0.73, 0.64)	**0.73**	(0.75, 0.72)
GMOD	0.84	(0.80, 0.87)	0.88	(0.84, 0.93)	**0.95**	(0.95, 0.95)
APP	0.69	(0.60, 0.82)	0.63	(0.53, 0.78)	**0.74**	(0.66, 0.83)
KON	**0.96**	(0.94, 0.99)	0.75	(0.78, 0.72)	0.90	(0.91, 0.90)
CJ	**0.95**	(0.94, 0.95)	0.74	(0.81, 0.68)	0.91	(0.93, 0.91)

Table 5: Testing on Law200Test

3.3 Integrating Hard Constraints

The CRF-based model presented in the previous section is limited in the sense that the prediction is based on the probabilistic linguistic information of current tokens and the contexts. Therefore, we integrate principle-based hard linguistic constraints into the probabilistic sequential model. To this end, we use the rule-based supertagger. The supertagger constitutes a set of grammar rules created based on a development set of Swiss law texts. It reduces morphosyntactic ambiguities, in particular, morphosyntactic case features, and assigns them to dependency grammar labels by applying grammar hand constraints and domain constraints. In the case of morphosyntactic and syntactic ambiguity, it assigns more than a morphosyntactic case feature and dependency label to tokens.

We use a feature-based method to integrate the hard constraints returned by the rule-based supertagger into a CRF model. In this combination method, the rule-based supertagger guides a CRF-based one in form of features. In this way, the morphosyntactic case features can be weighted and incorporated into a sequential CRF model as features, that is, parallel to other linguistic features. In the experiment, we added 300 sentences of law texts (Law300Training) with two additional features returned by the rule-based supertagger: (1) morphosyntactic case features and (2) the dependency labels. We trained CRF models on this in-domain training data combined with the adapted out-of-domain data (TüBa60NoLawNoise).

We tested this newly trained model on the test set of the law texts. The tagger was improved by the integration of the case features (cf. #F of tb. 4) and dependency relation labels (cf. #G of tb. 4). By combining these two features, the CRF model achieved the best improvement of all experiments (cf. #H of tb. 4).

3.4 Results

Table 5 shows that the rule-based tagger, the CRF-based

	Hybrid Supertagger		ParZu		Bohnet	
	0.90	(1824, 199)	0.88	(1783, 240)	0.87	(1771, 252)
SUBJ	0.88	(0.83, 0.94)	**0.90**	(0.87, 0.94)	0.87	(0.85, 0.89)
OBJA	**0.82**	(0.80, 0.84)	**0.82**	(0.77, 0.87)	0.77	(0.79, 0.75)
OBJD	**0.85**	(0.87, 0.84)	0.77	(0.77, 0.77)	0.69	(0.60, 0.82)
GMOD	**0.96**	(0.98, 0.93)	0.94	(0.95, 0.92)	0.95	(0.97, 0.92)
APP	**0.65**	(0.60, 0.70)	0.58	(0.52, 0.66)	0.59	(0.48, 0.75)
KON	**0.78**	(0.89, 0.69)	0.62	(0.50, 0.81)	0.62	(0.50, 0.83)
CJ	0.91	(0.94, 0.88)	0.93	(0.92, 0.95)	**0.94**	(0.94, 0.94)

Table 6: Evaluation

tagger (cf. #E of tb. 4), and the hybrid tagger (cf. #H of tb. 4) achieved label accuracies of 84%, 86% and 92%, respectively. With regard to the main dependency labels illustrated in tb. 5, the CRF-based tagger was improved by the guidance of the rule-based tagger. However, the hybrid tagger lowered the accuracy of the rule-based tagger with regard to coordination (`KON, CJ`).

We measure the strength of the integration of the CRF-based tagger (linguistic probability-based soft constraints) with the rule-based tagger (linguistic principle-based hard constraints) and its hybrid effects on performance, by calculating the degree to which the hybrid tagger was guided by the rule-based tagger. In the test data (Law200Test), 1512 tokens were nouns. The hybrid followed the guidance of the rule-based tagger on 1,286 tokens (85% of nouns), 97.51% of which were correctly tagged, that is, the label accuracy was 97.51%. The hybrid tagger accepted suggestions from one of the labels provided by the rule-based tagger on 127 tokens (8% of nouns). In this case, the label accuracy was 96.85%. Two systems did not agree and predicted different tags on 99 tokens (7% of nouns). In this case, the label accuracy of the rule-based tagger was 40.74%, whereas that of the hybrid tagger was 35.18%. Thus, the label accuracy of the hybrid CRF-based tagger was high if it was guided by the rule-based tagger, and it selected the outputs suggested by the rule-based tagger. Correspondingly, the label accuracy of the hybrid tagger was low if the CRF model did not accept suggestions from the rule-based tagger. In this case, both the rule-based tagger and the hybrid tagger performed poorly (label accuracy of 40.74% and 35.18%, respectively). These results indicated that the poor accuracy of the rule-based tagger led to the CRF's rejection of the suggestions. However, the alternative hybrid tagger was not very accurate (label accuracy of 35.18%).

3.5 Evaluation

We tested the hybrid CRF-based supertagger and two state-of-the-art parsers on 200 randomly selected and newly annotated sentences in Swiss law texts (Law200Eval). For this purpose, we use the ParZu parser [21] and the Bohnet parser [22]. ParZu is a dependency parser based on a rule-based component combined with statistical components. The Bohnet parser is a statistical parser that was the best parser for the labeling task in German in the CoNLL-2009 Shared Task. We trained the model on the training data of TüBa (TüBa60Train). Table 6 provides an overview of the results. Our hybrid supertagger outperformed the two parsers in label accuracy. The hybrid tagger achieved the best label accuracy (90%) and F1 score in main dependency grammar relations listed in Table 6: accusative and dative objects (`OBJA`, `OBJD`) and coordination (`KON`).

4. CONCLUSION

In this paper, we showed that linearly modelling hard and soft linguistic constraints is of relevance for assigning grammatical functions in German. The hybrid supertagger was able to boost the label accuracy on domain test data by adding a small amount of domain training data to a large amount of out-of-domain data that was adapted to the legislative domain. The in-domain training data was augmented with morphosyntactic case features and dependency labels provided by a rule-based tagger. This feature-based combination of the rule-based supertagger onto the CRF-based supertagger improved the label accuracy from 86% to 92%.

5. REFERENCES

[1] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proc. of the HLT-NAACL*, 2006.

[2] David McClosky and Eugene Charniak. Self-training for biomedical parsing. In *Proc. of the HLT*, 2008.

[3] Satoshi Sekine. The domain dependence of parsing. In *Proc. of the ANLP*, pages 96–102, 1997.

[4] Daniel Gildea. Corpus variation and parser performance. In *Proc. of the EMNLP*, pages 167–202, 2001.

[5] Rahul Goutam. Exploring self-training and co-training for Hindi dependency parsing using partial parses. In *Proc. of the IALP*, 2012.

[6] Rahul Goutam and Bharat Ambati. Exploring self-training and co-training for dependency parsing. In *Proc. of the CICLing*, 2012.

[7] Jennifer Foster, Joachim Wagner, Djamé Seddah, and Josef van Genabith. Adapting WSJ-trained parsers to the British National Corpus using in-domain self-training. In *Proc. of the IWPT*, pages 33–35, 2007.

[8] Roi Reichart and Ari Rappoport. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proc. of the ACL*, pages 616—623, 2007.

[9] Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. Bootstrapping statistical parsers from small datasets. In *Proc. of the EACL*, pages 331–338, 2003.

[10] Anoop Sarkar. Applying co-training methods to statistical parsing. In *Proc. of the NAACL*, pages 175–182, 2001.

[11] John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. In *Proc. of the ICML*.

[12] Tilman Höhle. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985s*, 1986.

[13] Claudia Kunze and Lothar Lemnitzer. *Computerlexikographie. Eine Einführung*. Güter Narr, Tübingen, 2007.

[14] Mariikka Haapalainen and Ari Majorin. GERTWOL: ein System zur automatischen Wortformerkennung deutscher Wörter. Technical report, Lingsoft, Inc., 1994.

[15] Peter F Brown, Peter V Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 1992.

[16] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2011.

[17] Daniel Jurafsky and James H Martin. *Speech and Language Processing*. Pearson Education International, New Jersey, 2009.

[18] Adam Pauls and Dan Klein. Faster and smaller n-gram language models. In *Proc. of the HLT*, volume 1, 2011.

[19] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proc. of the EMNLP*, pages 355–362, 2011.

[20] Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proc. of the ACL*, 2010.

[21] Rico Sennrich, Martin Volk, and Gerold Schneider. Exploiting synergies between open resources for German dependency parsing, POS-tagging, and morphological analysis. In *Proceeding of the RANLP*, pages 601–609, 2013.

[22] Bernd Bohnet. Very high accuracy and fast dependency parsing is not a contradiction. In *Proc. of COLING 2010*, 2010.