Identification of Comparable Argument-Head Relations in Parallel Corpora

Kathrin Spreyer, Jonas Kuhn and Bettina Schrader

Department of Linguistics
University of Potsdam
Germany
spreyer@uni-potsdam.de, kuhn@ling.uni-potsdam.de,
bschrade@uni-potsdam.de

Abstract

We present the machine learning framework that we are developing, in order to support explorative search for non-trivial linguistic configurations in low-density languages (languages with no or few NLP tools). The approach exploits advanced existing analysis tools for high-density languages and word-aligned multi-parallel corpora to bridge across languages. The goal is to find a methodology that minimizes the amount of human expert intervention needed, while producing high-quality search and annotation tools.

One of the main challenges is the susceptibility of a complex system combining various automatic analysis components to hard-to-control noise from a number of sources. In this paper, we present a series of systematic experiments investigating to what degree the noise issue can be overcome by (i) exploiting more than one perspective on the target language data by considering multiple translations in the parallel corpus, and (ii) using minimally supervised learning techniques such as co-training and self-training to take advantage of a larger pool of data for generalization. We observe that while (i) does help in the training individual machine learning models, a cyclic bootstrapping process seems to suffer too much from noise. A preliminary conclusion is that in a practical approach, one has to rely on a higher degree of supervision or spend some effort in the formulation of noise detection heuristics.

1. Introduction

Many phenomena of interest to syntactic, semantic and pragmatic research and high-level language technological development are relatively infrequent in running corpus data.[1] This means that a representative number of occurrences in corpora cannot be realistically hand-annotated using the standard methodology, i.e., designing annotation schemes and performing annotation of a corpus sample well in advance and independently of the application of these annotations. Moreover, carefully annotated corpora are only available for a small number of languages; and even when annotated corpora for different languages do exist, contrastive linguistic studies or multilingual linguistic engineering efforts are often complicated by differences in annotation schemes and/or in the types of genres sampled in the corpora. As a consequence, data-oriented language research often has to fall back to larger unannotated corpora. In practice, search on unannotated corpora is often based on ad hoc decisions, such as formulating queries with particular lexemes that are deemed to represent a whole class of items. This leads to an often tedious turnaround cycle of manually assessing search results and refining query expressions. Our hypothesis is that human effort can be channeled much more effectively if an interactive machine learning (ML) platform exploiting a combination of ideas and technologies is applied, such as (i) the annotation projection idea (Yarowsky et al., 2001), using (statistical) word alignments over the parallel corpus to transfer—or "project"—the analysis obtained by an existing tool for one language to the translational correspondence of the sentence in another language; (ii) progress in the development of parallel deep, but robust linguistic grammars for a number of languages (such as English, German, Japanese, as included in the grammars from the ParGram project (Butt et al., 2002)), which can be used as "hubs" in a multi-parallel corpus; (iii) machine learning techniques for combining various information sources; (iv) weakly supervised learning techniques for channeling human annotation effort.

The specific task we use to explore the methodology is the identification of a verb's arguments in one language in a parallel corpus (we are using Dutch in the experiments reported here), taking advantage of deep automatic analyses for two word-aligned translational correspondences (English and German), but not presupposing any tools for the target language.[2]

One of the main challenges for machine learning design in this context is that a complex system which combines various automatic analysis components is susceptible to hard-to-control noise from a number of sources. In this paper, we present systematic experiments exploring to what degree the noise issue can be overcome by (i) exploiting of more than one "view" on the target languages data, and (ii) using minimally supervised learning techniques such as co-training and self-training to take advantage of a larger pool of data for generalization.

Section 2. provides some background on the project context and briefly addresses related work; in Section 3. we present details of the architecture we adopt and the assumptions we make. Section 4. is a description of the specific data resources we used in our experiments, and Section 5. presents the experimental results we obtained. We end with a conclusion in Section 6.

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

[2]We do use the Alpino-parser for Dutch (Malouf and van Noord, 2004) in our experiments, but only as a convenient way of approximating a gold standard for development data and in order to simulate human expert annotation in approaches like Active Learning.

2. Context and Motivation

2.1. Project Context

The project is part of *Sonderforschungsbereich* 632 (SFB) on Information Structure at the University of Potsdam and Humboldt University Berlin, a long-term linguistic research network studying the linguistic realization of information-structural distinctions (categories like focus and topic) across languages. Here, as in many other scenarios, the exploration of large corpus resources with an open inventory of distinctions is a crucial step in a systematic study towards deeper understanding of tendencies in variable linguistic behaviour.

2.2. Related Work

The idea of exploiting parallel texts and crosslingual parallelism to transfer existing annotations in one language to a new language has first been brought forward by Yarowsky et al. (2001), who applied it to morphological analysis and NP bracketing. Their method of *annotation projection* has been adopted in a wide range of annotations, including part-of-speech tagging (Ozdowska, 2006), dependency parsing (Hwa et al., 2005) and role semantic analysis (Padó and Lapata (2006) for German, Padó and Pitel (2007) for French). However, these approaches differ from the one presented here in that we propose projection from two rather than a single language.

The redundant use of two sources of cross-lingual projection (L_1 and L_2) implements an old idea that has been discussed in various guises in quite different contexts (C1–C3).

(C1) Triangulation. In classical work on machine translation as well in recent work on statistical machine translation (Och and Ney, 2001; Cohn and Lapata, 2007), the idea of "triangulation" (originally due to Martin Kay) is considered a helpful tool for disambiguating translational choices: if some unit in language A can be translated to language B in several ways, there is a chance that an existing parallel translation to language C will disambiguate the choice in B. In the annotation projection scenario, where the word alignment, the parsers or translational mismatches are potential sources of error, triangulation can be directly applied to filter out less reliable target sentences.

(C2) Co-Training. In *co-training*, different (near-redundant) "views" on the same problem are used to train initial independent learners on a small labelled seed set and iteratively augment their training set with unlabelled data which (some of) the learners label most confidently (Blum and Mitchell, 1998). Parallel data can be divided into such views very naturally, with one view for each language (as Callison-Burch and Osborne (2003) did in statistical machine translation). Similar ideas can be exploited, involving a higher degree of expert intervention, as in corrected co-training (Hwa et al., 2003), self-training (McClosky et al., 2006) or active learning (Becker and Osborne, 2005).

(C3) ML techniques. Finally, general discriminative ML techniques like Maximum Entropy models or Support Vector Machines can deal with redundant feature information from alternative sources, i.e. in this case features based on the different projection sources and combinations thereof, such that the learner can exploit sources that have turned out to be reliable but can also back off to more general information when necessary.[3]

Our framework facilitates a combination of all three ways of exploiting parallel bases of projection, (C1) and (C2) in the selection of data for "partial" learners in a bootstrapping architecture, and (C3) as the central mechanism in learning.

3. A Platform for Multi-Source Annotation

This section outlines the components of the architecture proposed above. Section 3.1. describes the projection of annotations from multiple source languages, which is complemented in Section 3.2. by details about the representation of multilingual information as rich feature sets suitable for machine learning. Section 3.3. addresses the issue of noise which arises in a projection architecture, and how we intend to handle it.

3.1. Consensus Projection

The core idea of multi-source annotation projection is simple: we start from a large multi-parallel corpus of n languages (L_1-L_n), including two languages (L_1 and L_2), for which parallel parsers are available. The goal is to rapidly obtain reliable parallel annotations *for all n languages*, with annotations focusing on a particular targeted aspect. For each of the languages L_3 through L_n, this amounts to a combined projection and machine learning task, determining the correct target annotation for a surface string in that language (L_i), given word-aligned and parsed strings in languages L_1 and L_2.

Identification of argument-head relations. To develop and test the proposed general architecture, we define a specific target annotation task for experimentation: the identification of verb arguments in a language L_i included in a parallel corpus, with English and German as L_1 and L_2 and Dutch as L_i. This is illustrated in Figure 1. The task is for a given Dutch verb (here *stellen*) to identify those words in the Dutch string that are the lexical heads of the verb's arguments. Taking together all word-by-word argument decisions amounts to identifying the verb's argument frame.[4] The choice of this experimental task is motivated by a number of factors: usefulness for our broader project context, prototypicality of the task, and feasibility of detailed and independent evaluation. The task can be viewed as a generic example of frame labelling, i.e., any more specific template filling task is a variant in terms of the general procedure.

Training data are automatically annotated by means of *consensus projection* from L_1 and L_2 to L_i: heads and (the lexical heads of) their arguments are projected to L_i from both

[3] Parallelism in the grammatical analyses for L_1 and L_2 is more crucial for (C1) and (C2) than it is for (C3) (which could be applied with any combination of available tools), but since an understanding of the mechanics is crucial for tuning system performance (especially in a highly interactive architecture), it turns out immensely helpful even here that the analyses are comparable.

[4] In our current arguments, we do not distinguish argument labels, but this is a straightforward specialization of the task.

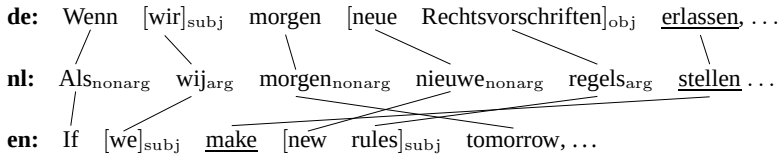

Figure 1: The argument (head) identification task.

the parse in L_1 and the one in L_2, using the word alignment. Where the L_1- and L_2-based projections coincide on the same word in L_i, we add this word (paired up with the verbal head) to the training data for the argument status classifier. If the projections are contradictory, i.e., there is a projection from one source language, but not the other, the word (or the entire sentence) is discarded for lack of evidence for either analysis.[5]

3.2. ML with Redundant Information

We pursue two goals in ML design, namely firstly to explore what architecture (decomposition, feature design etc.) is effective for modelling the target task and capable of exploiting multiple and potentially noisy views on the training data, and secondly to couch that ML architecture in a semi-supervised context in order to efficiently channel minimal human annotation effort.

To address the first goal, we adopt the general discriminative ML framework of Maximum Entropy (maxent) models, and focus on questions of feature design and decomposition of the complex task of assigning a set of arguments to a given verbal head. A natural decomposition assumes a language-based division of the feature space. Feature design can help in dealing with noise which may enter the data through the various component resources (automatic word alignment, source language parsers, true cross-language divergencies etc.). But the combination of diverse automatic tools of varying quality as information sources for complex high-level classification tasks places serious challenges on the machine learning approach taken. Hence we are couching our ML component in an active learning setting: We hypothesize that this issue can in part be resolved by adopting an interactive, semi-supervised approach in which the component tools represent different "views" on the relevant information (as the projection bases in different languages do in our setup). Such a sophisticated ML approach—with expert intervention as needed—may help to cancel out some of the noise.

On the basis of the projected argument classifications, we train maxent models using the MegaM software package (Daumé III, 2004). Our models rely on properties of the head h and the candidate argument w, with their aligned counterparts from L_1 and L_2. More specifically, the features we use to represent words fall into three categories:

Lexical features include surface form, lemma, and part-of-speech (POS) for all three languages, as well as NP form, tense, voice, aspect, verb type and clause type from the parallel grammars.

Alignment features encode information about the configuration of the automatic word alignment, namely the number of words that w aligns to (outdegree) and is aligned to (indegree), as well as an explicit binary indicator of bidirectionality (as an approximation of alignment confidence).

Contextual features are sentence length, position and distance relative to the head, and (for L_i) intervening complementizers, the number of intervening verbs, the presence of intervening (sentence-final) punctuation, POS of the preceding and following n-gram context (in the experiments, n was set to 2, resulting in the context window w_{j-2} w_{j-1} w_j w_{j+1} w_{j+2}). Moreover, contextual features record governing prepositions and whether w is part of a coordinate structure.

We also spell out (selected) feature conjunctions to capture interaction between features. A concrete example is given in Figure 2. It shows an excerpt from the feature vector generated for the word *regels* in Figure 1.

3.3. Dealing with Noise

Like any annotation projection framework, we encounter considerable noise in our data. The single most salient error source is the automatic word alignment, which is responsible not only for erroneous target annotations, but also for errors in the feature vectors, since we include features from the source languages depending on the word alignment.

In this paper, we investigate three techniques to address the problem of noisy training data: regularization, feature selection, and explicit representation of error sources. Regularization (or penalization) limits the relative weight that may be attributed to a single feature. In particular, we impose a Gaussian prior with mean $\mu = 0$ on the model parameters. The MegaM package provides an implementation of this mechanism, parametrized over inverse variance λ (Daumé III, 2004). Greater values of λ (and hence smaller variance) enforce feature weights closer to the mean of the prior.

Feature selection can be seen as a special case of regularization where some features are assigned zero weight (and hence excluded from the model altogether). Although feature selection is a more drastic technique than regularization, it is very attractive from a pragmatic perspective in that it effectively reduces the feature space and thus time and memory requirements in training and prediction. Here, we experiment with a very simple frequency threshold θ which discards all features that occur less than θ times in the training data.

[5] Discarding the entire sentence in the presence of discrepant projections implements an aggressive filter on the error-prone alignment information. It also has the advantage of counteracting further enforcement of the strong negative bias inherent to the data: less than 10% of all words in the test data are arguments.

	lexical	alignment	contextual
nl arg	postag=nounpl, lemma=regel	de-indeg=1, en-indeg=1	sentLen=16, position=pre prec1=pronpers, succ2=adv
nl head	postag=verbprespl, surface=stellen	de-indeg=1, de-outdeg=1	–
de arg	case=acc, postag=NN	bi=true, indeg=1	sentLen=17, position=pre
de head	pred=erlassen, vtype=main	bi=true, outdeg=1	–
en arg	ntype=common, num=pl	indeg=1, outdeg=1	position=post, dist=1
en head	pred=make, tense=pres	bi=true, outdeg=1	–

Figure 2: Multilingual feature vector (excerpt).

The third technique, the explicit representation of noise, is based on the intuition that meta-level information about the error source (here, the word alignment) might be used to identify configurations which give rise to inconsistent annotations. We attempt to achieve this by embedding alignment features (Section 3.2.) directly in the models.

Finally, we examine to what extent information about the true annotation for selected data points can improve performance even if the noise in the features remains. To limit the amount of labelled data that is required, a small set of Alpino-parsed sentences (cf. fn. 2) is used as seed data in two weakly supervised bootstrapping experiments (self-training and co-training). We also report on an experiment with corrected co-training (Hwa et al., 2003), an interactive bootstrapping method which combines co-training with ideas from active learning. In the self-training scenario, after training an initial classifier from the seed data, we iteratively add data points to the training set for which the model obtained in the previous iteration is most confident in classification (i.e., assigns a probability close to 0 or 1). Preliminary experimentation on a development set revealed that adding positive and negative examples in a proportion that approximates the empirical distribution observed in the seed set (less than 10% positives) severely impairs performance: recall drops as rapidly as precision increases. We therefore introduce two additional parameters p and n which control the proportion of, respectively, positives and negatives in the data points that are added in each iteration.

In the co-training scenarios, two near-redundant classifiers are derived to inform each other: one classifier is trained on the supervised seed annotations and features projected from German, the other one on the corresponding features from English. In contrast to self-training, the co-training criterion for selecting examples to be added to the training set is based on the agreement between the two classifiers. One would hope that the separation of projection sources enables correct treatment of examples that are ruled out by the consensus criterion (Section 3.1.), either because alignment links are missing or because of true cross-language divergence.

4. Data and Resources

Parallel corpus. For our experiments we use the parallel Europarl corpus (Koehn, 2005). It consists of translations of the proceedings of the European Parliament in 11 languages, each represented by approx. 1 million sentences (30 million words). The alignment on the word level was established using GIZA++ (Och and Ney, 2003), followed by lemmatization and POS tagging for the languages Dutch, German and English with the IMS TreeTagger (Schmid, 1994). We used the word alignment and a list of English main verbs extracted from the POS-tagged Europarl corpus to determine potential verbal heads in the target language. We thus reduce our reliance on the availability of a tagger for the target language.

Parallel grammars. The grammars we use to parse the German and English portion of the corpus are LFG grammars from the ParGram project (Butt et al., 2002), run in the XLE environment (Maxwell and Kaplan, 1991). In addition, we use XLE's built-in Prolog-based extraction engine to extract from the parses the features that we are interested in.

We should note that we experience considerable data reduction during preprocessing. Out of a subset of 300,000 sentences from Europarl, we found that a sentence alignment consistent across all three language pairs could be retrieved for only approx. 200,000 sentences. The parsers yielded a full parse in both languages for 81,000 of these sentences. In 8,500 sentence triples, the English sentence does not contain any of the verbs listed in our verb list, and for another 3,000 sentences the extraction engine fails. This results in a set of 69,500 sentence triples; the actual alignment of the selected heads (based on the verb list) to corresponding verbal elements in German and English was ultimately supported by the word alignment for almost 69,000 heads. Data loss of this magnitude may seem prohibitive at first, but it is relatively unproblematic given that the annotations involved so far are completely automatic, and hence no human resources are wasted. Moreover, we plan to integrate additional (shallower) tools that the projection mechanism can fall back to if deeper analysis fails, thus increasing recall.

Gold Standard. A small gold standard was annotated for the Dutch argument identification task. For a given sentence and verbal head, the heads of argument phrases are considered arguments. Argument status of phrases is based on a binarization of the guidelines of the spoken Dutch corpus CGN (Hoekstra et al., 2003). Annotation was carried out by a linguistically trained native speaker of Dutch. The gold standard consists of annotations for 240 verbal heads in 222 sentences, giving a total of 4,756 local datapoints.

	R	P	F
direct projections:			
German→Dutch	52.9	52.2	52.6
English→Dutch	48.8	54.3	51.4
consensus	34.1	74.6	46.8
German→Dutch w/ de+nl features	39.7	56.5	46.6
English→Dutch w/ en+nl features	32.8	59.5	42.3
consensus w/ de+en+nl features	**41.5**	**63.3**	**50.1**

Table 1: Evaluation of direct projection (top) and initial unsupervised models trained (2,000 sentences).

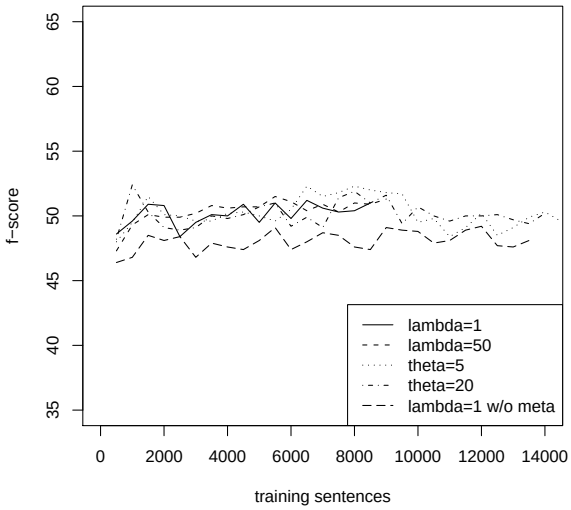

Figure 3: Effects of feature selection.

5. Experiments

On the basis of the feature vectors described in Section 3.2., we train an initial maxent model which is unsupervised in the sense that the training data was generated in a fully automatic way, without (simulated) human intervention. This model (the last line in Table 1) incorporates all available features, i.e. features about Dutch, German and English. For comparison, we also give results for the direct projections and for models that use as training classifications the annotations projected from a single source language, together with corresponding features from that language (and Dutch). The results in Table 1 confirm that the consensus-trained model with all features (f-score 50.1) indeed outperforms both single-source models (46.6/42.3).

It is interesting to note that, although single source projection fares well in terms of high recall, the maxent models trained on these data do not. As a result, the model trained on the consensus projection data outperforms the models trained on the single-source projection data in terms of precision *and* recall (both significant at p<0.01).

5.1. Feature Selection

Figure 3 shows the effect of the feature selection and regularization techniques (Section 3.3.) as the amount of training data (and hence the number of features) is increased. As expected, the learning curve of the unrestricted model ($\lambda = 1$) is nonmonotonic due to overfitting, while regularization with a narrow prior ($\lambda = 50$) is capable of stabilizing the performance.[6] However, the vast number of features causes the training algorithm to run out of memory when the training set size exceeds ∼9,000 sentences. The frequency filters on features obviate this shortcoming, but performance degrades with both the moderate ($\theta = 5$) and the aggressive ($\theta = 20$) threshold.

Figure 3 also illustrates the usefulness of the meta-level alignment features: the model trained without such information ('lambda=1 w/o meta') performs consistently worse than its informed counterpart. We attribute the success of the meta-information not to those features as such, but to their conjunction with other (lexical) features.

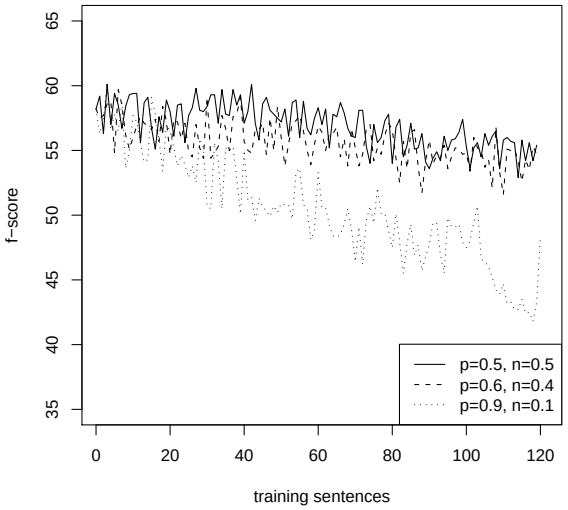

Figure 4: Self-training with selection bias.

5.2. Bootstrapping

As outlined in Section 3.3., our minimally supervised bootstrapping experiments address two scenarios, self-training and co-training. Both involve initial classifiers derived from 200 sentences parsed with the Alpino parser. At each subsequent iteration, we add to the training set 100 data points out of a pool of 100 sentences, based on a confidence (self-training) or agreement (co-training) measure. The pool of 100 sentences is sampled randomly from the entire pool of unlabelled sentences in each iteration.

Self-training. The f-score curves for the self-training experiments are given in Figure 4. While none of the classifiers actually improves, we observe that the selection bias introduced by the proportion parameters p and n can tackle the skewness problem to some extent: the curve for the classifier which adds positive and negative data points in a balanced proportion maintains a stable f-score, which means that recall is increased without overly inhibiting precision (not shown). The curve with $p = 0.9/n = 0.1$ illustrates that a stronger bias towards positive classifications amplifies this trend, which—like the inverse extreme—leads to unfavourable trade-offs between recall and precision and hence a drop in f-score.

Co-training. Figure 5 shows learning curves for 50 iterations of the co-training cycle. For each parameter setting,

[6] With $\lambda = 50$ we simultaneously increase the number of training iterations from 100 to 500 so as to allow convergence.

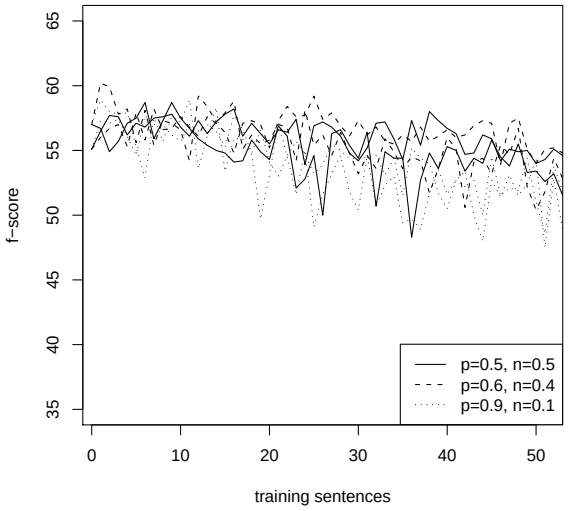

Figure 5: Co-training with selection bias.

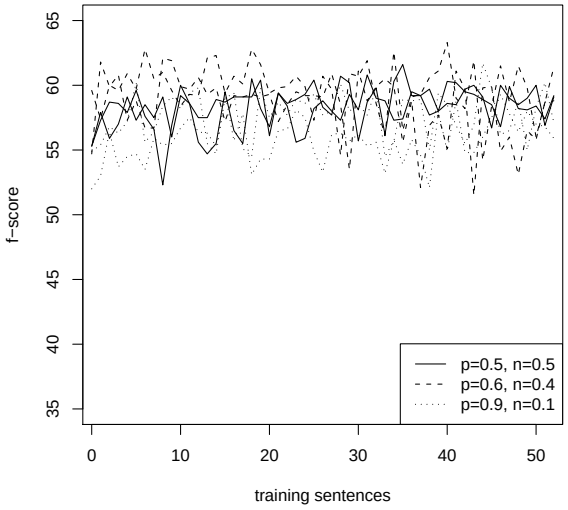

Figure 6: Corrected co-training with selection bias.

we give the performance of both classifiers (German- and English-based). Here, the impact of the selection bias is not as clear as in the self-training case: all settings produce highly unstable curves. In particular, corresponding classifiers appear to react to augmented training sets in opposite directions. Both the persistent discrepance between the paired classifiers and the limited usefulness (degrading performance) of the ensemble might be explained with reference to the results reported at the beginning of this section, namely that the consensus model with all features outperforms the single-source models. We tentatively conclude from this experiment that one source language in isolation is not sufficient to model the complex task when the training data contains noise in the classifications *and* in the features.

Corrected co-training. To test whether noise in the features has the same impact as noise in the classifications, and to ensure that manual annotations can improve performance at all if they are combined with faulty feature vectors, we impose stronger supervision on the co-training setup in that selected data points are added to the training set with their correct (Alpino-classified) label instead of the label predicted by the bootstrapped classifiers. The results in Figure 6 show that correct classifications can indeed compensate for errors in the feature vectors. The f-scores are slighly increasing as more data is seen, although the curves are still unsteady.

5.3. Discussion

Although none of the experiments affords a distinct improvement, they reveal important directions for further investigation. Regularization can be exploited to streamline the feature space and thereby cancel out inconsistencies and prevent overfitting. We expect that the combined application of a restrictive prior on the feature weights and the bootstrapping techniques can yield more monotonic changes in classification accuracy.

Furthermore, the weakly supervised models considered here exhibit performance gains of roughly 10% even with only 200 labelled sentences, which can be realistically annotated in less than a day.

6. Conclusion

We have described an interactive platform for multi-source projection in an ML context. Within this architecture, we have investigated the usefulness of several unsupervised and semi-supervised methods in the face of considerable noise in the training data. The experimental results reveal that all methods are seriously impaired in their effectiveness, albeit to varying degrees. Specifically, we have been able to show—in the unsupervised as well as the semi-supervised settings—that models which incorporate information projected from multiple source languages are more robust towards inconsistencies in the training data than the corresponding single-source models.

In future work, we will apply more sophisticated feature selection methods, in combination with general modular (statistical) heuristics to deal with noisy data.

In ongoing research (Bouma et al., 2008), we try to capture the interdependence of isolated argument status decisions explicitly in joint models and integrate our methodology with active learning proper to provide the linguistic researcher with an efficient interface to large parallel corpora.

7. References

M. Becker and M. Osborne. 2005. A two-stage method for active learning of statistical grammars. In *Proceedings of IJCAI 2005*.

A. Blum and T. Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 1998 Conference on Computational Learning Theory*, pages 92–100, July.

G. Bouma, J. Kuhn, B. Schrader, and K. Spreyer. 2008. A Framework for multi-source annotation projection. Ms., University of Potsdam, Germany.

M. Butt, H. Dyvik, T. Holloway King, H. Masuichi, and C. Rohrer. 2002. The parallel grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.

C. Callison-Burch and M. Osborne. 2003. Co-Training For Statistical Machine Translation. In *Proceedings of the 6th Annual CLUK Research Colloquium*.

T. Cohn and M. Lapata. 2007. Machine Translation by Triangulation: Making Effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 728–735, Prague.

H. Daumé III. 2004. Notes on CG and LM-BGFS optimization of logistic regression. Unpublished manuscript. Paper and software downloadable from `www.cs.utah.edu/~hal/megam/`.

H. Hoekstra, M. Moortgat, B. Renmans, M. Schouppe, I. Schuurman, and T. van der Wouden. 2003. Cgn syntactische annotatie. `http://ww2.tst.inl.nl/images/stories/docs/syn_prot.pdf`.

R. Hwa, M. Osborne, A. Sarkar, and M. Steedman. 2003. Corrected co-training for statistical parsers. In *Proceedings of the 20th International Conference on Machien learning*, Washington, D.C.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering*, 11(3):311–325.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit 2005*.

R. Malouf and G. van Noord. 2004. Wide coverage parsin with stochastic attribute value grammars. In *IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses*.

J. T. Maxwell and R. M. Kaplan. 1991. A Method for Disjunctive Constraint Satisfaction. In Masaru Tomita, editor, *Current Issues in Parsing Technology*, pages 173–190. Kluwer Academic, Boston, MA.

D. McClosky, E. Charniak, and M. Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 337–344.

F. J. Och and H. Ney. 2001. Statistical Multi-Source Translation. In *MT Summit 2001*, pages 253–258, Santiago de Compostela, Spain, September.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

S. Ozdowska. 2006. Projecting POS tags and syntactic dependencies from English and French to Polish in aligned corpora. In *Proceedings of the EACL 2006 Workshop on Cross-Language Knowledge Induction*, pages 53–60, Trento, Italy, April.

S. Padó and M. Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of COLING/ACL 2006*, Sydney, Australia.

S. Padó and G. Pitel. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. In *Proceedings of TALN-07*, Toulouse, France.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, England.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT 2001*.