Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection

Yannick Versley

SFB 833
University of Tübingen
E-mail: versley@sfs.uni-tuebingen.de

Abstract

We present work on tagging German discourse connectives using English training data and a German-English parallel corpus, and report first results towards a more comprehensive approach of doing annotation projection for explicit discourse relations.

Our results show that (i) an approach based on a dictionary of connectives currently has advantages over a simpler approach that uses word alignments without further linguistic information, but also that (ii) bootstrapping a connective dictionary using distribution-based heuristics on aligned bitexts seems to be a feasible and low-effort way of creating such a resource.

Our best method achieves an F-measure of 68.7% for the identification of discourse connectives without any German-language training data, which is a large improvement over a nontrivial baseline.

1 Introduction

Annotation projection is an approach based on using parallel text to transfer linguistic annotations from one language to the other (Bentivogli and Pianta, 2005; Pado and Lapata, 2005); using such techniques, it is possible to bootstrap automatic linguistic annotation for a particular purpose when the respective tools and/or resources are only available in another language – for example, Johansson and Nugues (2006) used such an approach to create a FrameNet parser for Swedish with only a bare minimum of hand-annotation.

In our case, the target consists in explicit discourse relations – discourse relations which are more easy to detect because of the use of so called *discourse connectives*. The category of discourse connectives, despite their common function of linking the contents of two different clauses, is syntactically heterogeneous: It includes coordinating and subordinating sentence conjunctions as the most prototypical examples, but also large and syntactically heterogeneous groups such as multi-word items with conjunction-like behaviour (*as soon as*, *as long as*), and single- or multi-word adverbials that show anaphoric, rather than syntactic, linking

behavior (e.g., *for example*, *in addition*, *on the contrary*). As discourse relations present an abstraction from the concrete (syntactic) means, we expect them to show little variability even in the case of translations that vary in surface word order or syntactic realization, making them an attractive target for annotation projection.

The Penn Discourse Treebank 2.0 (PDTB; Prasad et al., 2008) contains, for the text basis covered by the Wall Street Journal portion of the Penn Treebank, annotation of discourse relations marked by a connective (*Explicit*), those that are not marked by a connective (*AltLex* and *Implicit*), as well as annotations that do not signal a discourse relation (*EntRel* and *NoRel*). About half of the relations in the PDTB are in the *Explicit* category. In contrast to *Implicit* discourse relations, where even a very substantial annotated corpus such as the Penn Discourse Treebank is insufficient for training a reliable automatic classifier, previous research for English has established that finding and classifying explicit discourse relations robustly is well within the reach of the state of the art: Pitler and Nenkova (2009) report an accuracy of 95% for the disambiguation of connective versus non-connective readings of potential discourse connectives, and 94% for classifiying the signaled discourse relation into one of the four top-level categories in the PDTB's taxonomy. Even for the second level of the taxonomy (which is closer to the granularity level found in other discourse-annotated corpora), it is possible to classify instances with about 84% accuracy (Versley, 2011), which is close to the reported inter-annotator agreement for the corpus.

Even for explicit discourse relations, a sufficient amount of annotated data is necessary, as discourse connectives are often ambiguous (between discourse and non-discourse readings, or between different discourse relations), and because the set of discourse connectives is potentially large: The Penn Discourse Treebank contains slightly more than one hundred different discourse connectives; the German Handbuch der Konnektoren (Pasch et al., 2003)), a handbook describing the grammatical properties of German connectives, lists about 300 different connectives. The set of connectives is also syntactically (as well as semantically) heterogeneous, and is not necessarily limited to syntactic constituents. Hence, techniques to reduce the effort for annotating the necessary training examples would be very useful in the creation of discourse-annotated corpora for other languages.

While several medium-to-large discourse corpora exist for English (Carlson et al., 2003; Wolf and Gibson, 2005; Prasad et al., 2008), the availability of resources for other languages including German is much more limited. Among the existing resources for German, the handbook of Pasch et al. (2003) focuses on syntactic properties of different connectives (and would therefore need to be complemented with sense information). Two further resources, the lexicon DiMLex (Stede and Umbach, 1998) and a small RST-annotated corpus (Stede, 2004) have been described in the literature but are not publically available.

Using projection from automatically tagged instances on a parallel corpus, we tackle the problem of bootstrapping the annotation of discourse connectives by investigating (i) the variability in the translation of these items, and (ii) possible approaches to create an automatic tagger for German discourse connectives based on the annotated data.

In order to tag German text using the English training data, several intermediate steps are necessary: Firstly, the original training data has to be used to create automatic annotation for the English side of the parallel corpus (section 2); Second, the annotation on the English side of the parallel corpus has to be projected across the alignment to form training data for the German side of the corpus (section 3). Finally, the projected German data can be used to learn a classifier and annotate a gold-standard sample of German newspaper text (section 4).

2 Tagging of English Connectives

To tag connectives in English text, we use classifiers that are trained on data from the Penn Discourse Treebank 2.0 (PDTB; Prasad et al., 2008). We used an approach that allows fully automatic identification and disambiguation of discourse connectives that is loosely based on the work of Pitler and Nenkova (2009), with modifications that make it more useful for our task: Firstly, our approach creates tags that correspond to the finer second level of the PDTB's taxonomy of discourse relations. Secondly, Pitler et al make use of information that can be found in the hand-annotated treebank, but not in automatic parses (traces and semantic function labels such as '-PRP'), whereas our approach is able to reach similar accuracy (i.e., better by a fraction of a percent) using information that can be derived from automatic parses.

Using the 15 366 *Explicit* relations from sections 2-22 for training data, our tagger is able to distinguish between discourse and non-discourse usages of potential connectives with 92% precision and 98% recall (in cross-validation on the WSJ text); disambiguation of discourse instances between the four coarse relation types in the Penn Discourse Treebank (*Comparison*, *Expansion*, *Contingency* and *Temporal*) is possible with 94-95% accuracy, whereas distinguishing between the sixteen second-level relations (e.g., distinguishing between *Concession* and *Contrast*, *Cause* and *Condition*) is possible with about 84% accuracy. The third and finest level on the PDTB taxonomy can be disambiguated with 79% accuracy. These accuracy results mirror the decrease in annotator agreement reported by Prasad et al. (2008), which leads us to believe that they correspond to a greater difficulty in the disambiguation task (rather than widespread lack of features).

For the experiments, we automatically tagged the English side of the EuroParl corpus using the Berkeley parser[1] for syntactic preprocessing. On a sample of data from EuroParl corpus, we see that the use of automatic parsing and out-of-domain data leads to a slight decrease in performance, with 83% precision and 97% recall.

2.1 Syntax and Tense Features for English Connectives

Two of the features used in our discourse tagger are reimplementations of ones used by Pitler and Nenkova (2009): One is the string of the connective itself, with mod-

[1] http://code.google.com/p/berkeleyparser/

ifiers – such as "*two minutes*" in "*two minutes before the train departed*" omitted in order to ensure the generalizeability of the connective instances.

The second group of features comprises syntactic features, namely the labels of self, parent, left sibling and right sibling nodes (counting from the lowest node that covers all of the words annotated as connective span and that is not the only child of its parent), as well as additional features signaling the presence of a VP node or of a trace as a child of the right sibling.

A third group of features is based on arguments, which we can identify reliably for a restricted subset of the discourse connectives (subordinating and coordinating conjunctions, w-adverbials (*[S . . . [SBAR [WHADVP when] he sleeps]]*). For fronted (preposition- or adverb-headed) adverbials, we can reliably identify one of their arguments, which is the parent clause, sentence whereas the other argument is linked anaphorically and is not identifiably as easily.

Based on the identified arguments, we extract the following indicators:

- the part-of-speech of the first non-modal verb in the sentence (descending from the argument clause node into further VP and S nodes to cover both nesting of VPs and coordinated sentences)
- the presence (and word form) of modals and negation in the clause
- a tuple of *(have-form, be-form, head-POS, modal present)* as proposed by Miltsakaki et al. (2005).

(In the result tables, the part-of-speech/presence of modals pair of features will be called *pos/md*, whereas the tuple describing auxiliaries, the POS of the lexical head, and the presence of modals will be simply called *verb*).

Verb tense and modals are relatively shallow correlates of more interesting properties such as facticity or veridicality (i.e., whether the speaker asserts the propositional content of that clause to be true), but they are easy to extract in a robust manner and useful as a first approximation to a more comprehensive approach such as those of Palmer et al. (2007) to classifying situation entities.

3 Mapping Discourse Annotation

In order to create annotations on the German side, we have to project the automatically annotated data from the English side using available sentence and word alignments to create training data for a German classifier. In a simple first step, we can create a projected version of the English annotation by simply considering every token that has a word alignment link to an English connective;[2] in the case of discontinuous connectives (English *if/then* or *either/or*) or of discontinuous word alignments, the resulting connectives on the German side can be discontinuous.

[2]The word alignments themselves are postprocessed from the statistical alignments that are output by GIZA++, using the *intersection* or *grow-diag-final* heuristic. The experiments where the heuristic used is not stated use *grow-diag-final* since the intersective alignments are often too sparse.

Directly using the projected data on the German side can be problematic not just because of the noise from the statistical word alignment, but also due to a mismatch between the noisy alignments and the syntax-based mechanism used in the English connective finder: The approach chosen for English is dictionary-based in that a list of potential connectives is used to identify candidates by looking for occurrences of the particular word sequences, and subsequently using binary classification (based on syntactic and tense features) to filter out non-connective occurrences.

Dealing with the problem of noise in the projected annotation is possible in two different ways: One would be to deal with the problem by using a shallower sequence tagging approach in the connective classification that does not use a pre-established list of connectives; the other would be to use a pre-established list of connective candidates (i.e., word sequences that can have a connective function, but may be ambiguous between discourse and non-discourse readings), either from an external source such as the HdK, or induced by refining the connective candidates that can be extracted from the word-based projection.

3.1 A Simple CRF Baseline

As a baseline, we consider the most straightforward way to do annotation projection and learn a proposal mechanism: We project the discourse connective annotation on the English side using the word alignments (tagging all German words that are aligned to an English discourse connective), and use the tagging created through this method to train a sequence classifier.

Besides the words themselves, the sequence classifier uses features signaling the start and end of clauses (from automatic parses), and the types of those clauses.

This alignment approach creates relatively noisy annotation, as not all discourse connectives are present in the German translation. Witness the following example:

(1) *{Das}* *ist* *ganz* *im* *Sinne* *der* *Position,* *die* *wir* *als* *Parlament*
{That} is wholly in terms of the position, which we as Parliament
immer *vertreten* *haben.*
always advocated have.
[Indeed], it is quite in keeping with the positions this House has always adopted.

The English *Indeed*, which would signal that the sentence is an explanation of the previous sentence, is not present in the German translation. As a result, an arbitrary part of the sentence is aligned to the discourse connective and receives the connective span. The result is still useful, though, if the classifier that is learned somehow averages out the noise that occurs in training. The alternative to living with this noise, though, is to look for ways to improve the precision, as in the two following approaches.

3.2 A Dictionary-based Approach

An alternative to plain sequence tagging would be an approach more similar to the dictionary-based approach for English, where potential connectives are extracted using a (monolingual) list of such items and these candidates are filteres using a binary classifier (into actual discourse connectives, and word sequences that look like a discourse connective, but actually are not).

In the initial step, we use the list of connectives contained in the German *Handbuch der Konnektoren* to identify potential discourse connectives in the German EuroParl text; if multiple overlapping occurrences are found (e.g., *als*/*when* vs. *als ob*/*as if*), the longest match is kept.

To find out whether a given occurrence should be treated as a positive or as a negative example, we compare its span with all sets of words projected from potential connectives on the English side and use an overlap metric (Dice) to determine which potential connective string on the English side corresponds best. If the best match potential connective on the English side is tagged as a discourse connective, the German span is used as a positive example; if it is not aligned to a potential connective on the English side or the aligned string is tagged as not being a discourse connective, the German span is used as a negative example.

The subsequent binary classifier uses a language-independent version of the syntactic features that are used in the English-only classifier: the connective string, and features describing the lowest common node in the parse tree (label of self, parent, and left and right siblings). For the syntactic preprocessing of German trees, we use the parser of Versley and Rehbein (2009), with a grammar learned from the TüBa-D/Z treebank (Telljohann et al., 2009).

3.3 Inducing a Connective List

While the HdK provides us with a list of connectives, it is an interesting and potentially useful question whether we can induce such a list from the aligned data. As all word alignments have been created automatically, and translators occasionally omit or add discourse connectives in sentences, however, we have to correct or filter the word sequences that can be extracted from the alignments.

For each candidate string, we determine the following three statistics:

- the *total* number of occurrences
- the number of occurrences that overlap with a projected discourse connective (i.e., where at least one word of the candidate string is aligned to at least one word from the English discourse connective)
- for each aligned occurrence, a Dice-based overlap measure between the tagged English discourse connective and the projection of the candidate string (where 0 means no overlap and 1 means that they cover exactly the same words).

Using a dataset composed of the HdK list and a random sample of other candidate strings (both limited to those that had at least 15 aligned occurrences), we found out that the most effective method to discriminate between connectives and spurious candidates was to require a minimum average overlap of about 66-70% (over all aligned occurrences).

To build the list, we took all proposed strings that had at least 15 aligned occurrences, where the average overlap was at least 70% and where the product of (i) the average overlap and (ii) the ratio between aligned and unaligned occurrences was not smaller than $\frac{1}{25}$. These occurrences were then ordered by average overlap (considering better-overlappig proposed strings first) and discarding any proposed string where a subsequence had a higher average overlap.[3]

The resulting list contains 293 items, of which some are not in the HdK list, either as new connectives that fit the HdK criteria, or as items that would need to be manually corrected or filtered.[4]

4 Evaluation and Discussion

To evaluate performance on the German side, we annotated a text sample comprising slightly more than 5000 tokens of text from the TüBa-D/Z corpus, with two annotators independently performing the annotation and merging the differences, yielding 136 connective instances.

The annotated gold standard reflects the criteria set forth in the German *Handbuch der Konnektoren* for grammatical properties of a connective x:

- x cannot be inflected.
- x does not assign case to elements in its syntactic environment.
- x realizes a binary relation.
- The arguments of x are propositional.
- The arguments of x are clauses.

As can be seen in table 1, simply tagging every string from the HdK's list (*all HdK* as a discourse connective results in very good recall[5] but also poor precision.

[3] Keeping shorter proposed strings in the list does not change the end result much, since the tagging process will prefer longer matches over shorter ones.

[4] When ranked by average overlap, the first 62 candidate strings have a high proportion of connectives that are also part of the HdK (69%), some new items (18%, e.g. *anders ausgedrückt* in addition to *anders gesagt* as equivalent to *in other words*)), some are truncated (e.g., *facto* instead of *de facto*), or contain additional tokens such as commas or complementizers (10%), and some which do not fit the criteria for a discourse connective at all (3%). At the bottom of the list, the overlap with HdK items is substantially lower (29%), while the proportion of incomplete/longer items (35%) as well as incorrect items (19%) are much higher. The proportion of correct items not covered by the HdK stays about the same (17%).

[5] Note that the recall is not 100% since we found phrases that match the HdK's criteria for connectives, but are not part of the handbook's list.

	Prec	Recl	$F_{\beta=1}$
all HdK	27.0	**94.9**	42.1
simple CRF, giza-refined	74.0	41.9	**53.5**
simple CRF, giza-intersect	**83.9**	38.2	52.5
HdK+classifier	62.3	76.5	**68.7**
induced+classifier	58.3	56.6	57.5
HdK+CRF	74.7	43.4	54.9
induced+CRF	70.2	43.4	53.6

Figure 1: Evaluation results: Tagging German text (newspaper sample)

For the CRF approach, we used Léon Bottou's Stochastic Gradient Descent CRF learner[6] using default settings (50 training epochs, C=1.0). The CRF baseline yields a much better precision (both for the intersected alignments for *giza-intersect* and using the *grow-diag-final* heuristic for *giza-refined*) but relatively poor recall around 40%.

Using word alignments and the HdK word list we can also derive a binary classifier for occurrences of potential connectives. Such an approach gives a precision that is significantly better than the pure dictionary-based approach, with a comparably smaller loss in recall (76.5% against 94.9%, which is however still considerably better than the 41.9% reached by the CRF-based approach).

To establish whether the improvement in the dictionary-based approach is to be seen in the cleaner training data, or in the more expressive features that are used in the syntax-based classification, we performed additional experiments to reflect the utility of these modifications in isolation. One experiment uses the syntax-based approach with an induced lexicon instead of the HdK one (*induced+classifier*), which results in a substantial loss in comparison to the manually annotated list, but still visibly better results than for the CRF approach.

In contrast, using the CRF approach with training data derived in a different way – using a dictionary in addition to word alignments, and removing anything that cannot be mapped to an entry in the list – shows only very little improvement over the CRF-based method where raw projections were used.

4.1 Summary

In this paper, we presented an approach to transfer a tagger for English discourse connectives by annotation projection using a freely accessible list of connectives as the only German resource. Compared to the supervised approach of Dipper and Stede (2006), who reach 78% F-measure on positive instances for a selected sample of nine German connectives, our annotation projection approach fares reasonably

[6] http://leon.bottou.org/projects/sgd

well.

Ongoing work will concentrate on three main issues: One issue is to complement the annotation projection of discourse connectives with mechanisms to find their sense (i.e., the discourse relation they signal), as well as their arguments. While the mechanisms for argument finding as well as for sense disambiguation that are used for English should in principle also work with other languages, German annotation for these features is not available yet.

The second main issue consists in the word alignments we have used (heuristically refined results from GIZA++), which are admittedly geared towards use in machine translation rather than being optimized for linguistic quality. Since discourse connectives most often consist of function words (rather than content words, which are easier for unsupervised alignment), the alignment of discourse connectives is especially quality-sensitive. Quite possibly, using a more elaborate approach, such as the reordering approach of (Collins et al., 2005), or more comprehensive procedures, such as the direct alignment of parse nodes (Zhechev and Way, 2008; Tiedemann and Kotzé, 2009), can further improve the quality reached by the approach.

A third broad issue is the creation of more expressive features on the German side, including tense/mood-based features, which have been shown to be beneficial for English tagging of discourse connectives.

Acknowledgements The research reported in this paper was financed by the Deutsche Forschungsgemeinschaft (DFG) as part of Collaborative Research Centre (SFB) 833 "Constitution of Meaning". The author would like to thank Emily Jamison and the three anonymous reviewers for helpful comments on earlier versions of the paper and to Anna Gastel for performing part of the connective annotation.

References

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. *Natural Language Engineering*, 11(3):247–261.

Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current Directions in Discourse and Dialogue*. Kluwer.

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause reordering for statistical machine translation. In *ACL 2005*.

Dipper, S. and Stede, M. (2006). Disambiguating potential connectives. In *Proceedings of the Konvens-2006 Workshop on the Lexicon-Discourse Interface*.

Johansson, R. and Nugues, P. (2006). A FrameNet-based semantic role labeler for Swedish. In *Proceedings of Coling/ACL 2006*, pages 436–443, Sydney, Australia.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A., and Webber, B. (2005). Experiments on sense annotations and sense disambiguation of discourse connectives. In *TLT 2005*.

Pado, S. and Lapata, M. (2005). Cross-lingual projection of role-semantic information. In *Proceedings of HLT/EMNLP 2005*.

Palmer, A., Ponvert, E., Baldridge, J., and Smith, C. (2007). A sequencing model for situation entity classification. In *ACL 2007*.

Pasch, R., Braußе, U., Breindl, E., and Waßner, U. H. (2003). *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin / New York.

Pitler, E. and Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *ACL 2009 short papers*.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Stede, M. (2004). The Potsdam Commentary Corpus. In *ACL'04 Workshop on Discourse Annotation*.

Stede, M. and Umbach, C. (1998). DiMLex: A lexicon of discourse markers for text generation and understanding. In *Coling 1998*.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., and Beck, K. (2009). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

Tiedemann, J. and Kotzé, G. (2009). Building a large machine-aligned parallel treebank. In *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories (TLT'08)*.

Versley, Y. (2011). Towards finer-grained tagging of discourse relations. In *Beyond Semantics: Corpus-based investigations of pragmatic and discourse phenomena (Workshop at the annual meeting of the DGfS)*. to appear.

Versley, Y. and Rehbein, I. (2009). Scalable discriminative parsing for German. In *Proc. IWPT 2009*.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Zhechev, V. and Way, A. (2008). Automatic generation of parallel treebanks. In *Coling 2008*.