Combining Parallel Treebanks and Geo-Tagging

Martin Volk, Anne Göhring, Torsten Marek
University of Zurich, Institute of Computational Linguistics
volk@cl.uzh.ch

Abstract

This paper describes a new kind of semantic annotation in parallel treebanks. We build French-German parallel treebanks of mountaineering reports, a text genre that abounds with geographical names which we classify and ground with reference to a large gazetteer of Swiss toponyms. We discuss the challenges in obtaining a high recall and precision in automatic grounding, and sketch how we represent the grounding information in our treebank.

1 Introduction

Treebanks have become valuable resources in natural language processing as training corpora for natural language parsers, as repositories for linguistic research, or as evaluation corpora for different NLP systems. We define a treebank as a collection of syntactically annotated sentences. The annotation can vary from constituent to dependency or tecto-grammatical structures. The term treebank is mostly used to denote manually checked collections, but recently it has been extended to also refer to automatically parsed corpora.

We have built manually checked treebanks for various text genres (see section 3): economy texts, a popular science philosophy novel, and technical user manuals. We are now entering a new genre, mountaineering reports, with the goal to link textual to spatial information. We build French and German treebanks of translated texts from the Swiss Alpine Club. This genre contains a multitude of geographical names (e.g. mountains and valleys, glaciers and rivers). Therefore we need to include the identification and grounding of these toponyms as part of the annotation process.

In this paper we first describe our corpus of alpine texts, then our work on creating parallel treebanks which includes aligning the parallel trees on word and phrase level. We sketch the difficulties in disambiguating the toponyms and describe our integration of the toponym identifiers as a special kind of semantic annotation in the treebank.

2 Our Text+Berg Corpus

In our project Text+Berg[1] we digitize alpine heritage literature from various European countries. Currently our group digitizes all yearbooks of the Swiss Alpine Club (SAC) from 1864 until today. Each yearbook consists of 300 to 600 pages and contains reports on mountain expeditions, culture of mountain peoples, as well as the flora, fauna and geology of the mountains.

The corpus preparation presented interesting challenges in automatic OCR correction, language identification, and text structure recognition which we have described in (Volk et al., 2010).

As of March 2010 we have scanned and OCR-converted 142 books from 1864 to 1982, corresponding to nearly 70,000 pages. This resulted in a multilingual corpus of 6101 articles in German, 2659 in French, 155 in Italian, 13 in Romansch, and 3 in Swiss-German. The parallel part of our corpus currently contains 701 translated articles amounting to 2.6 million tokens in French and 2.3 million tokens in German.

3 Parallel Treebanks

In recent years the combined research on treebanks and parallel corpora has led to parallel treebanks. We have built a parallel treebank (English, German, Swedish) which contains 1500 sentences in three languages: 500 sentences each from Jostein Gaarder's novel "Sophie's World", from economy texts (e.g. business reports from mechanical engineering company ABB and from the bank SEB), and from a technical manual with

[1] See www.textberg.ch.

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 192–196,
Uppsala, Sweden, 15-16 July 2010.

usage instructions for a DVD player (Göhring, 2009).

We have annotated the English sentences according to the well-established Penn Treebank guidelines. For German we followed the TIGER annotation guidelines, and we adapted these guidelines also for Swedish (see (Volk and Samuelsson, 2004)). For French treebanking we are looking for inspiration from the Le Monde treebank (Abeillé et al., 2003) and from L'Arboratoire (Bick, 2010). The Le Monde treebank is a constituent structure treebank partially annotated with functional labels. L'Arboratoire is based on constraint grammar analysis but can also output constituent trees.

3.1 Our Tree Alignment Tool

After finishing the monolingual trees we aligned them on the word level and phrase level. For this purpose we have developed the **TreeAligner** (Lundborg et al., 2007). This program comes with a graphical user interface to insert or modify alignments between pairs of syntax trees.[2]

The TreeAligner displays tree pairs with the trees in mirror orientation (one top-up and one top-down). This has the advantage that the alignment lines cross fewer parts of the lower tree. Figure 1 shows an example of a tree pair with alignment lines. The lines denote translation equivalence. Both trees are constituent structure trees, but the edge labels contain function labels (like subject, object, attribute) which can be used to easily convert the trees to dependency structures (cf. (Marek et al., 2009)).

Recently we have extended the TreeAligner's functionality from being solely an alignment tool to also being a powerful **search tool over parallel treebanks** (Volk et al., 2007; Marek et al., 2008). This enables our annotators to improve the alignment quality by cross-checking previous alignments. This functionality makes the TreeAligner also attractive to a wider user base (e.g. linguists, translation scientists) who are interested in searching rather than building parallel treebanks.

3.2 Similar Treebanking Projects

Parallel treebanks have evolved into an active research field in the last decade. Cmejrek et al. (2003) have built a parallel treebank for the specific purpose of machine translation, the Czech-English Penn Treebank with tecto-grammatical dependency trees. Other parallel treebank projects include Croco (Hansen-Schirra et al., 2006) which is aimed at building a English-German treebank for translation studies, LinES an English-Swedish parallel treebank (Ahrenberg, 2007), and the English-French HomeCentre treebank (Hearne and Way, 2006), a hand-crafted parallel treebank consisting of 810 sentence pairs from a Xerox printer manual.

Some researchers have tried to exploit parallel treebanks for example-based or statistical machine translation (Tinsley et al., 2009). Since manually created treebanks are too small for this purpose, various researchers have worked on automatically parsing and aligning parallel treebanks. Zhechev (2009) and Tiedemann and Kotzé (2009) have presented methods for automatic cross-language phrase alignment.

There have been various attempts to enrich treebanks with semantic information. For example, the Propbank project has assigned semantic roles to Penn treebank sentences (Kingsbury et al., 2002). Likewise the SALSA project has added frame-semantic annotations on top of syntax trees from the German TIGER treebank (Burchardt et al., 2006). Frame-semantics was extended to parallel treebanks by (Padó, 2007) and (Volk and Samuelsson, 2007). To our knowledge a treebank with grounded toponym information has not been created yet.

4 Geo-Tagging

Named entity recognition is an important aspect of information extraction. But it has also been recognized as important for the access to heritage data.

In a previous project we have investigated methods for named entity recognition in newspaper texts (Volk and Clematide, 2001). In that work we had only distinguished two types of geographical names: city names and country names. This was sufficient for texts that dealt mostly with facts like a company being located in a certain country or having started business in a certain city. In contrast to that, our alpine text corpus deals with much more fine-grained location information: mountains and valleys, glaciers and climbing routes, cabins and hotels, rivers and lakes. In fact the description of movements (e.g. in moun-

[2] The TreeAligner has been implemented in Python by Joakim Lundborg and Torsten Marek and is freely available at `http://kitt.cl.uzh.ch/kitt/treealigner`.

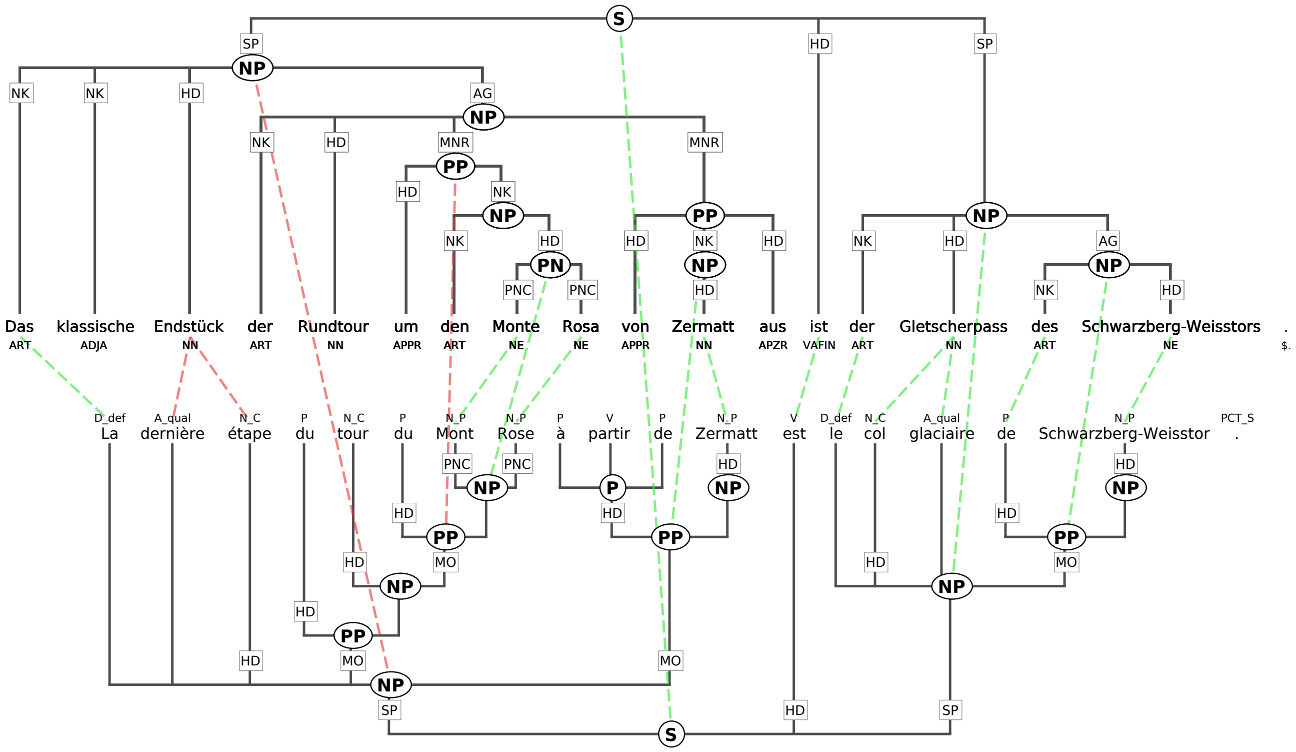
S
SP
NP
HD
AG
NK
MNR
PP
PN
PNC
Das klassische Endstück der Rundtour um den Monte Rosa von Zermatt aus ist der Gletscherpass des Schwarzberg-Weisstors .
ART ADJA NN ART NN APPR ART NE NE APPR NN APZR VAFIN ART NN ART NE $.
D_def A_qual N_C P N_C P N_P N_P P V P N_P V D_def N_C A_qual P N_P PCT_S
La dernière étape du tour du Mont Rose à partir de Zermatt est le col glaciaire de Schwarzberg-Weisstor .
P
MO

Figure 1: German-French tree pair with alignments in the TreeAligner.

tains) requires all kinds of intricate references to positions and directions in three dimensions.

In order to recognize the geographical names in our corpus we have acquired a large list of Swiss toponyms.

4.1 The SwissTopo Name List

The Swiss Federal Office of Topography (`www.swisstopo.ch`) maintains a database of all names that appear on its topographical maps. We have obtained a copy of this database which contains 156,755 names in 61 categories. Categories include settlements (10 categories ranging from large cities to single houses), bodies of water (13 categories from major rivers to ponds and wells), mountains (7 categories from mountain ranges to small hills), valleys, mountain passes, streets and man-made facilities (e.g. bridges and tunnels), and single objects like hotels, mountain cabins, monuments etc. Some objects are subclassified according to size. For example, cities are subdivided into main, large, middle and small cities according to their number of inhabitants.

Every name is listed in the SwissTopo database with its coordinates, its altitude (if applicable and available), the administrative unit to which it belongs (usually the name of a nearby town), and the canton.

4.2 A First Experiment: Finding Mountain Names

We selected an article from the SAC yearbook of 1900 to check the precision and recall of automatically identifying mountain names based on the SwissTopo name list. The article is titled “Bergfahrten im Clubgebiet (von Dr. A. Walker)”. It is an article in German with a wealth of French mountain names since the author reports about his hikes in the French speaking part of Switzerland. We took the article after OCR without any further manual correction. After our tokenization (incl. the splitting of punctuation symbols) it consisted of 9380 tokens.

We used the SwissTopo mountain names classified as “Massiv, HGipfel, GGipfel, and KGipfel” i.e. the 4 highest mountain classes. They consist of 5588 mountain names. This leads to a recall of 54 mountain names (20 different mountain names) at the expense of erroneously marking 6 nouns *Gendarm, Haupt, Kamm, Stand, Stein, Turm* as mountain names.

How many mountain names have we missed to identify? A manual inspection showed that there are another 92 mountain names (35 different mountain names) missing. So recall of the naive exact matching is below 40% despite the large gazetteer. We have reported on a number of reasons for missed names in (Volk et al., 2010).

We found that spelling variations and partial coreferences account for the majority of recall problems. In addition we need to disambiguate between name-noun and name-name homographs. This leaves the issue on how to represent the geotagging information in our treebank.

5 Geonames in Treebanks

Named entity classification can be divided into name recognition, disambiguation and grounding. The first two steps are applicable to all kinds of names. The final step of grounding the names is different depending on the name types. A person name may be grounded by refering to the person's Wikipedia page. The same could be done for a geographical name. The obvious disadvantage are changing URLs and missing Wikipedia pages. The goal of grounding must be to link the name to the most stable and most reliable "ground". Therefore toponyms are often linked to their geographical coordinates. We have chosen to link the toponyms from our alpine texts to unique identifiers in the SwissTopo database. This works well for Swiss names and particularly well for parallel French-German sentence pairs. The cross-language alignment assures that the names are recognized in either language and the classification information can then automatically be transfered to the other language.

In our example in figure 1, the mountain name "Monte Rosa" is listed in SwissTopo with its altitude (4633 m) and its location close to Zermatt. Since "Zermatt" itself occurs in the sentence, this is strong evidence that we have identified the correct mountain, and we will attach its SwissTopo identification number in our treebank. Technically this means we add a reference to the gazetteer and to the identifier within the gazetteer into the XML representation of the linguistic object.

In our German example sentence "Monte Rosa" is annotated as a proper name (PN). This occurrence is phrase 502 in sentence 311 of our treebank. The grounding id (g_id) is taken from SwissTopo which then allows us to access the geographical coordinates, the altitude and neighborhood information.

```
<nt id="s311_502"
    cat="PN"
    g_source="SwissTopo"
    g_id="7355873" >
```

Instead of integrating the grounding pointers directly in the XML file of the treebank, it is possible to use stand-off annotation by connecting the identifier of the geo-name with the identifier from the gazetteer in a separate file.

The alignments in our parallel treebank lead to the advantage that the grounding information needs to be saved only once. In our example, the corresponding mountain name "Mont Rose" in the French translation is listed in SwissTopo only as a building in the municipality of Genthod in the canton Geneva. Since we have strong evidence from the German sentence, we can rule out this option.

Zermatt itself occurs in both the French and German sentences in our example. It is listed in SwissTopo with its altitude (1616 m) and classified as mid-sized municipality (2000 to 10,000 inhabitants). Zermatt is a unique name in SwissTopo and therefore is grounded via its SwissTopo identifier. Likewise we ground "Schwarzberg Weisstor" (spelled without hyphen in SwissTopo) which is listed as foot pass in the municipality of Saas-Almagell. In case of doubt we could verify that Saas-Almagell and Zermatt are neighboring towns, which indeed they are.

6 Conclusions

Grounding toponyms in parallel treebanks represents a new kind of semantic annotation. We have sketched the issues in automatic toponym classification and disambiguation. We are working on a French-German parallel treebank of alpine texts which contain a multitude of toponyms that describe way-points on climbing or hiking routes but also panorama views. We are interested in identifying all toponyms in order to enable treebank access via geographical maps. In the future we want to automatically compute and display climbing routes from the textual descriptions. The annotated treebank will then serve as a gold standard for the evaluation of the automatic geo-tagging.

References

Anne Abeillé, Lionel Clément, and Francois Toussenel. 2003. Building a Treebank for French. In Anne Abeillé, editor, *Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*, chapter 10, pages 165–187. Kluwer, Dordrecht.

Lars Ahrenberg. 2007. LinES: An English-Swedish parallel treebank. In *Proc. of Nodalida*, Tartu.

Eckhard Bick. 2010. FrAG, a hybrid constraint grammar parser for French. In *Proceedings of LREC*, Malta.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006. The SALSA corpus: A German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, pages 969–974, Genoa.

Martin Cmejrek, Jan Curin, and Jiri Havelka. 2003. Treebanks in machine translation. In *Proc. Of the 2nd Workshop on Treebanks and Linguistic Theories*, pages 209–212, Växjö.

Anne Göhring. 2009. Spanish expansion of a parallel treebank. Lizentiatsarbeit, University of Zurich.

Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 2006. Multi-dimensional annotation and alignment in an English-German translation corpus. In *Proceedings of the EACL Workshop on Multidimensional Markup in Natural Language Processing (NLPXML-2006)*, pages 35– 42, Trento.

Mary Hearne and Andy Way. 2006. Disambiguation strategies for data-oriented translation. In *Proceedings of the 11th Conference of the European Association for Machine Translation (EAMT)*, pages 59–68, Oslo.

P. Kingsbury, M. Palmer, and M. Marcus. 2002. Adding semantic annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference (HLT'02)*, San Diego.

Joakim Lundborg, Torsten Marek, Maël Mettler, and Martin Volk. 2007. Using the Stockholm TreeAligner. In *Proc. of The 6th Workshop on Treebanks and Linguistic Theories*, Bergen, December.

Torsten Marek, Joakim Lundborg, and Martin Volk. 2008. Extending the TIGER query language with universal quantification. In *Proceeding of KONVENS*, pages 3–14, Berlin.

Torsten Marek, Gerold Schneider, and Martin Volk. 2009. A framework for constituent-dependency conversion. In *Proceedings of the 8th Workshop on Treebanks and Linguistic Theories*, Milano, December.

Sebastian Padó. 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Saarland University, Saarbrücken.

Jörg Tiedemann and Gideon Kotzé. 2009. Building a large machine-aligned parallel treebank. In *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories*, pages 197–208, Milano.

John Tinsley, Mary Hearne, and Andy Way. 2009. Exploiting parallel treebanks to improve phrase-based statistical machine translation. In *Computational Linguistics and Intelligent Text Processing*. Springer.

Martin Volk and Simon Clematide. 2001. Learn-filter-apply-forget. Mixed approaches to named entity recognition. In Ana M. Moreno and Reind P. van de Riet, editors, *Applications of Natural Language for Information Systems. Proc. of 6th International Workshop NLDB'01*, volume P-3 of *Lecture Notes in Informatics (LNI) - Proceedings*, pages 153–163, Madrid.

Martin Volk and Yvonne Samuelsson. 2004. Bootstrapping parallel treebanks. In *Proc. of Workshop on Linguistically Interpreted Corpora (LINC) at COLING*, Geneva.

Martin Volk and Yvonne Samuelsson. 2007. Frame-semantic annotation on a parallel treebank. In *Proc. of Nodalida Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages*, Tartu.

Martin Volk, Joakim Lundborg, and Maël Mettler. 2007. A search tool for parallel treebanks. In *Proc. of Workshop on Linguistic Annotation at ACL*, pages 85–92, Prague.

Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef. 2010. Challenges in building a multilingual alpine heritage corpus. In *Proceedings of LREC*, Malta.

Ventsislav Zhechev. 2009. *Automatic Generation of Parallel Treebanks. An Efficient Unsupervised System*. Ph.D. thesis, School of Computing at Dublin City University.