Extracting spatial information: grounding, classifying and linking spatial expressions

[Extended Abstract]

Frank Schilder*
R & D
Thomson Legal & Regulatory
610 Opperman Drive
Eagan, MN 55123, U.S.A.

Yannick Versley
AB WSV
Department for Informatics
Vogt-Kölln-Str. 30
22527 Hamburg, Germany

Christopher Habel
AB WSV
Department for Informatics
Vogt-Kölln-Str. 30
22527 Hamburg, Germany

ABSTRACT

This paper is concerned with the tagging of *spatial* expressions in German newspaper articles, assigning a meaning to the expression and classifying the usages of the spatial expression and linking the derived referent to an event description.

In our system, we implemented the activation of concepts in a very simple fashion, a concept is activated once (with a cost depending on the item that activated it) and is left activated thereafter. As an example, a city also activates the nodes for the region and the country it is part of, so that cities from one country are chosen over cities from different countries.

A test corpus of 12 German newspaper articles was tested regarding several disambiguation strategies. Disambiguation was carried out via a beam search to find an approximately cost-optimal solution for the conflict set of potential grounding candidates for the tagged spatial expression. Test showed that the disambiguation strategies improved accuracy significantly.

1. INTRODUCTION

in this paper we focus on the extraction and in particular the derivation of the meaning of spatial expressions. Some previous work has been carried out wrt. the task of named entity recognition (NER) , but the NER task does not derive the *meaning* of the expressions recognized. In particular, no disambiguation is carried out. The city *Paris*, for example, may refer to the French capital or to a town in Texas, USA. Moreover, *Paris* can also be used in a metaphorical sense as in (1):

(1) Paris rejected the "logic of ultimatums."

In (1), *Paris* refers to the French Government and not to the geographical entity of the French capital.

Consequently, the extraction task at hand is more complicated than tasks carried out within TREC[1] or machine learning competitions such as the Conference on Natural Language Learning (CoNLL).[2] Moreover, the extraction and disambiguation of spatial expressions is also embedded within a further extraction task: the spatial-temporal anchoring of events. Like the effort to anchor event description wrt. the time line [1, 2], we try to link geographical and geo-political entities to the event found in the same clause.

The presented system carries out the information extraction and linking process using the following steps:

1. Extraction of spatial expression candidates
2. Derivation and disambiguation of meaning
3. Resolution of anaphoric expressions
4. Linking to event descriptions

2. THE SPATIAL TAGGER

When tagging the names of places or geopolitical entities (i.e. the political body which is in control of some place), grounding is a useful prerequisite for further reasoning. The grounding task is carried out with the help of gazetteers, as described in the following section. In section 2.2, we describe the tagging system and its components in more detail before an evaluation of the system is presented in section 2.3.

2.1 Gazetteers used

For countries, place names and regions we used the ISO-3166-1 identifier, the UN-ECE locode identifier, [3] , and ISO 3166-2 identifiers, respectively. Other things such as

*Most of the work was carried out while this author was still at the University of Hamburg.

[1] http://trec.nist.gov/
[2] http://cnts.uia.ac.be/conll2003/

continents and political superstructures are handled on an ad-hoc basis.

2.2 Tagging system

The extraction of spatial information is done using a non-deterministic top-down-parser. The full paper will provide a brief introduction to the overall system architecture.

In our system, we implemented the activation of concepts in a very simple fashion, a concept is activated once (with a cost depending on the item that activated it) and is left activated thereafter. As an example, a city also activates the nodes for the region and the country it is part of, so that cities from one country are chosen over cities from different countries.

The actual process of disambiguation consists of identifying sets of conflicting possibilities and then finding an optimal solution (in term of activation costs when processed in normal reading order) for choosing one member of each set. In the current version of the disambiguator, we progress through the text in normal reading order while pruning less likely alternatives (beam search) yielding approximately linear cost complexity in terms of both search time and search space.

2.2.1 *Classification of the expressions*

After the place names have been identified and disambiguated, the role of the place name is determined using several rules. As an example, a spatial expression is used as the name of a location if it is preceded by a spatial preposition.

Since metonymous usage of location names is quite frequent in news texts, we have to distinguish possibly metonymous from non-metonymous (locational) use. The following distinctions are made (corresponding to the `type` attribute of the SpatexML description):

- locational use:
  the place name refers to a geographical region (in which events could have taken place).
- governmental body:
  the place name is used metonymously and designates the government of a country or a region, as in "Paris rejected the 'logic of ultimatums' ".
- attributive use:
  an adjective (or a noun) that somehow refers to a place name, but does not fall into the first two categories. Possibilities include:
  - an attribute to a relational noun:
    the French king, the Italian ambassador, the German government. Or, the king of France, the government of Germany.
  - the nationality, place of birth, living place or cultural affiliation of a person: the Turkish baker
  - the (former) headquarter or most important selling market of a corporation: the Dresdner Bank, British Airlines, Washington Post
  - the origin or destination of transported goods: the Korean missile parts

  There does not need to be such a relation, as there are also French fries, flying Dutchmen etc.

We also detect commonly occurring patterns as in the German government, the French capital and resolve these using a (limited) knowledge base.

2.3 Evaluation

The spatial tagger was evaluated with respect to the basic NER task of recognizing locations (`SPATEX (NE)`), the grounding of spatial expressions without (`SPATEX (ISO)`) and with their classification regarding different usages (`SPATEX (ISO, CLASS)`). 12 German newspaper articles were collected from internet news sites, tagged with POS information and fed to the spatial tagger system. We then pre-annotated the texts with an early version of the system, corrected the errors and added annotations for the expressions the system was unable to handle (for example, anaphoric references as in *das Land/the country*).

The following base lines were tested:

Base No disambiguation was carried out. The spatial expressions were extracted and classified solely based on gazetteer and linguistic information.

POS Location names were disambiguated against first names, proper names and closed-class lexemes using information from the Part-Of-Speech tagger.

IDS Disambiguation of location names using iterative deepening, yielding a global optimum with respect to the cost assigned.

Since the results with beam search do not differ visibly (less than 1%) from those where iterative deepening search was used, we omitted them in the table.

ANA Resolution of anaphora, using the last mentioned name of the matching class as referent. Since we model anaphora and disambiguation of location names as parts of the same process, we cannot evaluate the anaphora resolution separately from the disambiguation.

As can be seen below, the reliance on the Part-of-Speech tagger has a small cost in terms of the recall value (0.66 vs. 0.63) when city and country names get classified as normal nouns, but immensely helps the precision (0.60 vs. 0.95).

Even if we are only interested in locational references and not in the other classes, the distinction between classes is helpful. If we tagged everything as locational use in the absence of contextual hints, the precision for non-metonymous use would fall from 80% to 51%, while recall would be improved only slightly from 51% to 60%.

The linking of spatial expressions and event descriptions was also tested. This task, however, is quite easy, if the tagging of spatial expressions and event expressions was carried out correctly. Fulfilling this task depends mainly on the event

	Base	POS	IDS	POS+IDS	**POS+IDS+ANA**
`SPATEX (NE)`					
Precision	0.60	0.95	0.60	0.95	**0.94**
Recall	0.66	0.63	0.66	0.63	**0.74**
F-value	0.63	0.76	0.63	0.76	**0.83**
`SPATEX (ISO)`					
Precision	0.51	0.82	0.53	0.85	**0.84**
Recall	0.57	0.55	0.59	0.57	**0.66**
F-value	0.54	0.66	0.56	0.68	**0.73**
`SPATEX (ISO+CLASS)`					
Precision	0.45	0.73	0.46	0.75	**0.74**
Recall	0.50	0.49	0.51	0.50	**0.58**
F-value	0.47	0.59	0.48	0.60	**0.65**

Table 1: Evaluation results

tagging task. Since we did not focus on this task in this paper, we did not carry out a further evaluation of the tagging of the *SPATLINK* tags.

3. PROBLEMS

I the full paper, we will discuss a number of problems one has to face when spatial expressions are grounded.

4. CONCLUSIONS AND FUTURE WORK

In this paper we showed how spatial expressions can be extracted, grounded, classified and linked to event descriptions. We used a POS-tagger and a collection of gazetteers to tag the spatial expressions in news articles. The disambiguation of the extracted spatial descriptions was carried out by finding an approximately cost-optimal solution for the disambiguation problem. The classification of the spatial description with respect to three different usages (locational usage, governmental body and attribute use) was done employing linguistic clues from the context (e.g. spatial prepositions). The evaluation shows how the accuracy of the system improves when different disambiguation techniques are used.

5. REFERENCES

[1] E. Filatova and E. Hovy. Assigning time-stamps to event-clauses. In *proceedings of ACL'01 workshop on temporal and spatial information processing*, pages 88–95, Toulouse, France, 2001.

[2] I. Mani, B. Schiffman, and J. Zhang. A robust retrieval engine for proximal and structural search. In *Proceedings of Human Language Technology Conference and the conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, Edmonton, Canada, 2003.

[3] UNECE. LOCODE-Code for Trade and Transport Locations. Technical Report and UNECE Recommentation 16, United Nations Economic Commission for Europe, 1998.