Winter Semester 2009/10
Prof. Dr. Detmar Meurers
Holger Wunsch

ISCL, Universität Tübingen
Computational Approaches
to Functional Elements

Lab Session November 17, 2009:
A Majority Baseline for Predicting German Determiners and Articles

1. The TüBa-D/Z and TüBa-D/S corpora

We will work with the two German corpora TüBa-D/**Z** (**Z**: "Zeitung") and TüBa-D/**S** (**S**: spoken). TüBa-D/Z is a collection of manually annotated newspaper articles from the daily newspaper "die tageszeitung (taz)". TüBa-D/S are transcribed spoken dialogs about negotiating appointments such as flights, meetings, and so on. TüBa-D/S is manually annotated as well. Both treebanks share the same annotation principles and use the same POS tagset.

`/afs/sfs/lehre/dm/ws-09-10-functional-elements/corpora/tuebadz-4.tt`
(TüBa-D/Z)

`/afs/sfs/lehre/dm/ws-09-10-functional-elements/corpora/tuebads.tt`
(TüBa-D/S)

The format of these files is the same as the BNC files.

2. The POS tagset of the two treebanks

The POS tagset used in TüBa-D/Z and TüBa-D/S is the "Stuttgart-Tübingen Tagset" (STTS). It is documented at

`http://www.ifi.uzh.ch/~siclemat/man/SchillerTeufel99STTS.pdf`

3. Extending the majority baseline predictor to German

The goal of this sub-project is to extend our running systems for predicting definite articles and prepositions using a majority baseline approach to German.

1. Adapt your program such that you can determine the frequency of definite and indefinite articles and prepositions in the German treebanks.
2. Determine frequencies of articles and prepositions separately for TüBa-D/Z and TüBa-D/S.
3. Compare and discuss the frequency distributions.
4. With respect to the definite and indefinite articles, there is a significant difference between English and German. What is this difference? What impact does it have on our prediction strategy based on a majority baseline?

4. Three-stage evaluation for definite and indefinite articles

In the English version of the majority baseline predictor, we determined the correctness of a prediction simply by comparing the surface forms of the predicted element and the target (why did this work well?).

In German, articles agree with the nominals they determine in case, number, and gender, which adds complexity to the question when a prediction is correct. To get a feeling for the effect of the inflectional properties of articles on the performance of the system, we research it in three steps:

1. determine the correctness of the prediction of definiteness and indefiniteness, *ignoring* morphological variation in form.

2. evaluate using surface string comparison. This is much stricter than (1), since it takes into account the agreement properties. However, many articles are ambiguous between multiple forms (*die*: nom/acc sg fem, nom/acc pl masc/fem/neut). These underlying agreement properties are not taken into account using this evaluation strategy.

3. predict a surface form **and** its explicit morphological properties. Thus, instead of just the ambiguous *der* (which can be nom sg masc, gen/dat sg fem, acc sg masc, gen pl masc/fem/neut), you predict the unambiguous combination *der + dat sg fem*.

   The evaluation must then take into account both the surface form and the inflectional properties.

For steps 1 and 2, please reuse your existing prediction system.

Step 3 will require some modifications to the evaluation component. Furthermore, note that only TüBa-D/Z contains a morphological annotation layer, therefore you can only use TüBa-D/Z in step 3, but not TüBa-D/S.

You can find a version of the TüBa-D/Z with an additional third column containing the morphological annotation at

`/afs/sfs/lehre/dm/ws-09-10-functional-elements/corpora/tuebadz-4-morph.tt`