11-712: NLP Lab Report

Lingpeng Kong

April 26, 2013

Abstract

This is an incomplete draft of the development report for YAGMA (Yet Another German Morphological Analyser). The main content of this report includes basic information about German, past work on morphology of German, a survey of German morphology, design plans and experiment results for YAGMA.

1 Basic Information about German

German (Deutsch) is spoken as a first language by approximately 100 million people in the Federal Republic of Germany, the German Democratic Republic, Austria, Switzerland, and elsewhere, and as a second language by many others in Central and Eastern Europe. (Fox, 1990)

Most German vocabulary is derived from the Germanic branch of the Indo-European language family (European Commission, 2004). A number of words are derived from Latin and Greek, and fewer from French and English. German is written using the Latin alphabet. In addition to the 26 standard letters, German has three vowels with umlauts (Ä/ä, Ö/ö, and Ü/ü) and the letter ß.

In terms of morphological system, German has a particularly complex word structure, rather more complex than some other languages, including English. (Fox, 1990) It has a relatively speaking complex case and gender system, and very complicated rules in conjugation, declension and word formation.

2 Past Work on the Morphology of German

A lot of work has been done in the study of German morphology. Traditional linguistics provides a descriptive summary of the morphological phenomena in German (Boase-Beier, 2003) (Fox, 1990), which gives us insights of what is morphology in German, and how methods like "Umlaut" been used in word formation.

In the computational linguistics side, a handful of papers have been published (Lezius, 1996) (Lezius, 1998) (Lezius, 2000) (Schmid, 2004). These efforts fall into mainly different categories. The first one is to use a traditional stemming-then-generation approach, the basic idea of which is to cut all the possible affix and umlaut variations, then use the possible generation method to recover all the possible forms of the stem, if one (or more) of them match(es) the word given, we know its morphological components. The second approach is based on Finite State Transducer (FST), like much of the modern morphological analysers. These tools differs from each other mainly due to their coverage of the morphological phenomena (some only consider inflections while others only consider derivations). SMOR (Schmid, 2004) is a very comprehensive morphological analyser which consider

both inflections and derivations. It also make rules to deal with simple compounding phenomena in German and achieve a good result.

Despite the difficulties in German, it seems that it can still be solved in FST framework. Not much work (if any) tries to use upgrade to the context free grammar.

3 Available Resources

There are many available resources that can contribute to German morphological analysis tasks.

- Negra[1] and TIGER[2] are most widely used two Treebank for German. TIGER has morphological labels for the words since version 2.1.
- The CoNLL-2003 shared task data files[3] contain some labelled German text, where it give the words, lemma and its POS tags.
- Morphisto[4] is a morphological analyzer and generator for German wordforms. The basis of Morphisto is the open-source SMOR morphology for the German language developed by the University of Stuttgart (GPL v2) for which a free lexicon is provided under the Creative Commons 3.0 BY-SA Non-Commercial license.
- Vollformenlexikon[5] provides a German full form dictionary with about 90,000 base forms and 431,000 full forms (inflected forms). Each full form comes with full morphological information. The data is originally extract from Morphy (Lezius, 1996).
- LDC also provides several useful German corpus. (LDC95T11[6] is used for test for Schmid (2004))
- SFST - Stuttgart Finite State Transducer Tools [7]

In terms of the choice of our corpus A, B, C, we decided to use the TIGER Corpus since its labels are relatively high quality. We may also consider to mix some of the full forms in Vollformenlexikon because the TIGER labels are unique (words' morphological roles in a sentence are unique but only given full forms, we should have multiple possibilities for each word) while Vollformenlexikon labels all the possible morphological analysis of single full form word.

4 Survey of Phenomena in German

To make life easier, instead of talking countless exceptions and irregular rules, we will first introduce the most productive rules in German. Of course, this set of rules fall into the inflection category. These rules is the mainly focus of our work. After that, we will briefly discuss other morphological phenomena in German.

[1] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html
[2] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
[3] http://www.cnts.ua.ac.be/conll2003/ner/
[4] http://code.google.com/p/morphisto/
[5] http://www.danielnaber.de/morphologie/index_en.html
[6] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC95T11
[7] http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html

4.1 Case and Gender

German has a relatively complex case and gender system. It has four cases, the NOMINATIVE, ACCUSATIVE, GENITIVE, and DATIVE. It has three genders: MASCULINE, FEMININE, and NEUTER (in the plural, no distinction of gender is made).

4.2 Productive Morphological Phenomena (Inflection)

4.2.1 Nouns

Nouns can add some affix to generate their plural forms[8].
N [+SUFFIX (e)n, nen] → Plural
N [+SUFFIX s] → Plural
N [+SUFFIX e] ([+UMLAUT]) → Plural

Nouns can also have variations in cases.
Strong:
SIN-N [+SUFFIX (e)s] → Genitive
PLU-N [+SUFFIX (e)n] → Dative
Others remain unchanged.
Weak:
N [+SUFFIX (e)n] → All cases except Nominative

4.2.2 Verbs

Verb stems can add tense affix to generate infinite, past, past participle form.
Weak (Main Part of the German):
$[stem]_v$ + [+SUFFIX (e)te] → Past
[+PREFIX ge] + $[stem]_v$ + [+SUFFIX (e)t] → Past Participle

4.2.3 Adjectives

Adjectives can be various in cases, genders, and numbers[9].
Weak:
$[stem]_{adj}$ + [+SUFFIX e] → SIN MAS/NEU/FEM NOMINATIVE
$[stem]_{adj}$ + [+SUFFIX e] → SIN NEU ACCUSATIVE
$[stem]_{adj}$ + [+SUFFIX en] → Others

In German, adjectives also have their comparative and superlative forms. This is quite like in English.
$[stem]_{adj}$ + [+SUFFIX er] → Comparative
$[stem]_{adj}$ + [+SUFFIX (e)st] → Superlative

[8] A full discuss of these rules, like the choice of the rule based on different conditions will not be listed here to save space. Do note these rules exist, for example, when the noun ending with el, we should use the first rule rather than the second and third

[9] More precisely, when used with nouns in different cases, genders, and numbers, the adjectives have different forms.

4.3 Less Productive Morphological Phenomena

German has many exceptional and complicated rules (otherwise it will not considered to be thus difficult), but we won't worry too much about the irregular morphological changes, after all, they are finite, which means in worst cases, we can deal them with a table list all of them.

4.3.1 Derivations

Derivational, or lexical, morphology differs from inflectional morphology in being concerned with the formation of different lexemes from the same root, rather than with different grammatical forms of the same lexeme. Derivation is also very important in German, but they are less productive. Same thing happens in English. Some main phenomena related to derivation is summarized here[10].
[PREFIX ent/er/...] + $[stem]_v \rightarrow$ Verb
$[stem]_v$ + [SUFFIX eln/ern/...] $\rightarrow$ Verb
$[stem]_v$ + [SUFFIX er/...] $\rightarrow$ Noun
$[stem]_v$ + [SUFFIX bar] $\rightarrow$ Adjective
$[stem]_n$ + [SUFFIX ig] $\rightarrow$ Adjective
$[stem]_n$ + [SUFFIX isch] $\rightarrow$ Adverb

Conversion (Zero-derivation) also happens a lot in German, which always forms a word of a different category from the base word. But it is not really productive (Boase-Beier, 2003). Therefore, it seems that remember all of them in the finite state transducer is a reasonable solution for this.

4.3.2 Compounding

Compounding is one of the main difference between German and English. In German, compounding is every where. VV structure is quite rare, but we can still see many forms like NN NA AA AN and PN. This compounding may not only have the direct concatenate of two words, something may also be added in this process, like the following example:
$[stem]_n$ + s + $[stem]_n \rightarrow$ Noun
Fortunately, the good thing here, is that, after compounding, same rule set can be applied to it, the compounding noun has no particular features, it just performs as a common noun. Cases are also true for verbs and others.

4.4 Discussions and Notes

Despite the common affix things, German has some phenomena which worth more attentions. First is that, some morphemes are discontinuous, like the ge - en and ge - t. Second it that, a morpheme can be realized by suffix and umlaut together. These facts contribute to the complexity of the finite state transducer for German morphology analysis.

Other phenomena also do appear in German, like Suppletion, Ablaut, Chipping and Reduplication. We will not go through each of them here.

[10] Again, we do not tend to list all. We want to provide a general intuition about what happens in German.

5 Initial Design

We design the morphological analyser strictly follows the traditional FST design path. The big picture is like this: We first keep all the lexicon we care about in the lexicon file. Then, we represent every morphological phenomena as a special "suffix" to the lexicon (we will talk about this in details in later sections). After that, we use rules to process those special "suffix" and do some FOMA tricks to make the insertion, deletion, umlaut thing really happen in the string. Finally, we will delete all the intermediate symbols we inserted in the representation and make it the final surface string.

5.1 Empirical Word Cluster

Rather than focusing on the morphological rules and way too many exceptions in German, our approach, like the in Schmid (2004), is to put words into empirical word clusters based on what we should do to them to transform them into other morphological forms. For example, we have a class "NFem_0_n" which stands for Fem Nouns which remain the same (0) in the single 4 different forms (Nom, Acc, Dat, Gen) and have a suffix "n" in the plural 4 different forms. By doing this, we avoid most difficulties in hand coding rules. We manually check more than 50 classes for nouns using the German morphology dictionary. The lexicons come from Morphisto, current number of nouns we covered in the lexicon file is around 10,000.

5.2 Suffix Phenomena Solution

The suffix thing is really easy to handle. Like for English, our approach is to map the tags (e.g. +Fem, +NN etc.) directly to the morphology boundary and the suffix. The rule in FOMA is like the following:

```
+Fem+NN+Nom+Plural:^s #
```

where we map the +FEM+NN+Nom+Plural tags to "s". You can guess this is a rule for the word class who uses the "s" as the Plural suffix.

5.3 Deletion Phenomena Solution

For the deletion problem, unlike in English (e-deletion), we do not have a clear rule to do the deletion and the characters deleted are different from each other and often more than one character. Therefore, we use the different deletion rules to solve the problem, for example, if we want to delete the *a* in the end of the word, we implement an "Adel" (A-Deletion rule) to deal with that. Basically, if we first add a "-a" string before the morphological boundary (in the lexicon file), and then we use the rule to delete all the "a-a" before the morphological boundary (in the rule file).

```
define Delus {us} -> 0 || ?* _ "-us" "+um"* "^" ?* ;
```

While German has many kinds of deletions (-us, -um, -a etc.), the simply e-deletion rule in English does not work so well again. By applying this approach, we actually allowing different deletion rules based on the multi-character symbol ("-us" "-a"), and more importantly, this avoid the conflicts with the umlaut rules – while it is impossible for a single word has two deletion rules at the same time, the umlaut and the deletion can actually happen in the same stage. Here, by putting the deletions rule symbols before the umlaut symbols, we managed them together.

5.4 Prefix Phenomena Solution

The prefix phenomena is solved using a trick FOMA regular expression which stands for a insertion before the longest possible match. When we want to insert the "ge" at the beginning of the word (and "t" in the end, which is simple), we first put a special symbol "+ge" just after the morphology boundary. So in the second stage, when we see that, we use the following rules to insert it at the beginning of the string.

```
define Addpref   .?* "^" "+ge" ?*. @-> {ge} ...;
```

where the "@− >" stands for the left to right longest match in FOMA, and "..." represent the original string waiting for insertion.[11]

5.5 Umlaut Phenomena Solution

The umlaut phenomena are regarded as the most difficult ones in German. But it can also be solved with the approach we described above (although not perfectly). The rules for umlaut is like the following:

```
define Plusum1 a -> "+uma" || C* _ RC*  RuleSymbolPM* "+um" "^" ?* ;
```

Where we first map the character into intermediate symbols like the "+uma", and then map it into the umlaut characters like "ä". By performing the trick, we avoiding the errors which happens if there are two different vowels in the same word.
Also, an open question is that, when there are two vowels in the same word, it is not clear for which one we should do the umlaut thing (It is neither always the first vowel nor the last vowel). Since we found that the vowel u seems to be more preferable in umlaut, we insert a tricky rule which replace the u first than a or o.

6 System Analysis on Corpus A

The system analysis is performed use the gold morphology labels from the Tiger Corpus. We first extract 2000 nouns (and 2000 verbs) from the corpus with gold morphology label (1000 for corpus A and 1000 for corpus B), and then use them as input for our morphology analyser. Since in the Tiger corpus, words have their context, while for our morphology analyser, the decision is independent of context, we say we predict the word right if anyone of our predictions match the gold morphology label. The following table shows the coverage and the accuracy for corpus A for both nouns and verbs. (Regarding to verbs, we only consider the regular verbs and the "inf" or "past-participial" forms, so the accuracy is perfect but the coverage is actually quite limited.)

	Predicted	Corrected	Accuracy
Corpus A	573	560	0.978

Table 1: Nouns in Corpus A

[11] The rule is firstly not like this, we will talk about it in Lessons Learned and Revised Design.

	Predicted	Corrected	Accuracy
Corpus A	195	195	1.00

Table 2: Verbs in Corpus A

7 Lessons Learned and Revised Design

The first thing we found in Corpus A version of the analyser is actually a bug. We firstly write the umlaut rule as the following:

```
define Addpref   .?* "^" "+ge" ?*. @-> ge ...;
```

which perform well in most cases, but when the noun contains "ge" itself, it just can't recognise it. (The FST view, and the words looks all right, it just mysteriously goes wrong, I am wondering if it is a bug of FOMA.) But by adding the {} around the "ge", this bug fixed.

The other issue we learnt from the performance of corpus A is that the coverage is not good mainly due to the compounding words in German. So we add some rules to deal with the compounding words. We covered two types of the compounding rules, the first type has a hyphen inside the words, where we just simply ignore the first part of the word and regard all the morphological transformations happen in the second part of the words. The second type we consider is a base word followed by another base word, like the NN compounding words we mentioned in the previous section. The solution for this is also similar to the first type, where we add all the possible lexicons before the original grammar we have, so that we know the morphological transformations also happen in the second part of the word, which is consistent with the phenomena we observed in German.

8 System Analysis on Corpus B

By applying the methods mentioned in the previous section, we can see a substantial improvement in the coverage, where we cover 10% more words than before and do not really drop a lot in accuracy.

	Predicted	Corrected	Accuracy
Corpus B	650	615	0.946

Table 3: Nouns in Corpus B

	Predicted	Corrected	Accuracy
Corpus B	142	142	1.00

Table 4: Verbs in Corpus B

Here we also shows the revised solution for the original corpus A for reference.

	Predicted	Corrected	Accuracy
Corpus A	657	631	0.960

Table 5: Nouns in Corpus A

9 Future Work

Since German has very rich morphological phenomena, we still have a very long way to go before it is a reasonable German morphology analyser. The most important thing is:

- Continuing hacking on the word clusters and getting these done for verbs and adjectives also.
- Implementing more sophisticated rules for the compounding words, a reasonably well-preformed guesser should be needed in this step.
- Revising the rules for umlaut, to deal with the umlaut phenomena where there are multiple vowels in the same word.

References

Jean Boase-Beier and Ken R Lodge. *The German language: A linguistic introduction.* Wiley-Blackwell, 2008.

European Commission. Many tongues, one family. Languages in the European Union., 2004.

Anthony Fox. *The structure of German.* Oxford University Press, USA, 1990.

Wolfgang Lezius. Morphologiesystem morphy. In *Linguistische Verifikation, Dokumentation zur ersten Morpholympics 94*, pages 25–35, 1996.

Wolfgang Lezius. Morphy-german morphology, part-of-speech tagging and applications. In *Proceedings of the 9th EURALEX International Congress*, pages 619–623, 2000.

Wolfgang Lezius, Reinhard Rapp, and Manfred Wettler. A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for german. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 743–748. Association for Computational Linguistics, 1998.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. Smor: A german computational morphology covering derivation, composition and inflection. In *Proceedings of LREC 2004*, pages 1263–1266. Citeseer, 2004.