

Evaluation of the *BananaSplit* Compound Splitter

Niels Ott

March 6, 2006

1 Introduction

This document describes the evaluation process of the *BananaSplit* compound splitter which is a part of the *BananaRelation* tool. The task of the latter is to measure semantic relatedness of German compounds using GermaNet.

The compound splitter was extended in several points. This document describes only points that are new in respect to the original project presentation. (See `Compound-GermaNet-Slides.pdf`.)

The procedure of evaluation for the compound splitter is based on work of Köhn and Knight (2003). The authors describe a compound splitter used for machine translation and its evaluation procedure. The *BananaSplit* tool is based on Langer (1998).

2 The Output of *BananaSplit*

The compound splitter creates a simple analysis of the input word and outputs a flat bracketing structure. An important point is the purpose of using GermaNet: Words that exist as whole entries in GermaNet are not required to be split anyways. Therefore some compounds are not split on purpose. Some examples:

Garageneinfahrt

```
[.N [.N Garage ] [.B -Ø+n ] [.N Einfahrt ] [.U -Ø+Ø ] [.I -Ø+Ø ] ]
```

Aktionspläne

```
[.N [.N Aktionsplan ] [.U -a+ä ] [.I -Ø+e ] ]
```

The first example shows how bounding suffixes are handled: They are attached to a *B*-node. The second example demonstrates umlauting (*U*) and inflection (*I*) and at the same time shows that some compounds do not need to be split.

3 Simplifications

Before coming to actual evaluation, some concepts need to be simplified:

1. Restriction to nouns, verbs, and adjectives: The relatedness measures used by *BananaRelation* can handle only those three POS, so the gold standard should contain only these.
2. Drop structure: The bracketing output is in many ways eye, the interesting point is the correct splitting.
3. Drop parts-of-speech: As *BananaSplit* picks the first matching word sense from GermaNet only, POS is not really a hard criterion concerning the output of the program.

These points lead to a drastically simplified format of the output (compare with the data in section 2):

Garageneinfahrt
Garage Einfahrt

Aktionspläne
Aktionsplan

Note that the second example was changed to its base form but still not split as the word *Aktionsplan* is present as a whole in GermaNet.

4 Gold Standard

4.1 TIGER Sampler

The gold standard was taken from the TIGER Sampler that comes along with TIGER-Search. All verbs, nouns, and adjectives were extracted. From those, a random sample of 150 words was taken.

4.2 Annotation

The gold standard was annotated according to what would be the desired output as given in section 3. The limitations of the project do not allow the specification of a complex annotation scheme etc.

5 Metrics and Measuring

5.1 Procedure

The output of *BananaSplit* was converted to the simplified format specified in section 3. Output differing from the gold standard was checked manually.

Related to Köhn and Knight (2003), there are a couple of cases that can occur. They need to be adapted to my scenario:

- **Correct split:** Exact match with the gold standard. No manual checking.
- **Correct not A:** Words that should not be split and were not. No manual checking.

- **Correct not B:** Words that are split in the gold standard but not by the program because they are present as a whole in GermaNet. Manually checked.
- **Correct ambiguity:** In some cases, gold standard and program annotation differ but both are correct. E.g. *beschäftigten* could have both *beschäftigter* and *beschäftigt* as lemmata¹. Manually checked.
- **Wrong faulty split:** Words that should be split and were split, but in the wrong way. Manually checked.
- **Wrong split:** Words that should not be split but were split. Manually checked.
- **Lexicon failure:** Words that could not be analyzed at all because they are missing in GermaNet. No manual checking.

The lexicon failures are of course cases of wrong splitting. Yet I drop them for a simple reason: The relatedness measurement of *BananaRelation* would not be able to handle them anyways. Due to technical reasons it is not possible to separate the automatically checked cases of Correct Split and Correct Not A. Lexicon Failures are identified by the program.

5.2 Baseline

Maybe the results are not that good? As a baseline I take the words with no processing (raw). All words (including the Lexicon Failures) lead to 93 errors in 150, which yields an accuracy of 38%.

This is not a fair comparison for *BananaSplit* as I dropped Lexicon Failures. So with excluding the words that caused Lexicon Failures, the raw baseline is 78 errors in 119 words, resulting in an accuracy of 34.45%.

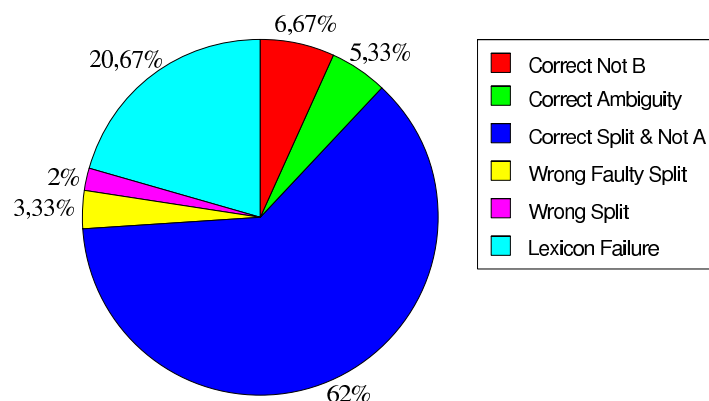
5.3 Upper Bound

The limitations of the evaluation experiment do not involve more than one annotator so agreement of annotators is not available. Therefore, no upper bound is available.

6 Results

Case	Occurrences
Correct Not B	10
Correct Ambiguity	8
Correct Split & Not A	93
Wrong Faulty Split	5
Wrong Split	3
Lexicon Failure	31
Sum	150

¹Consider that spelling resp. case is not distinguished so no disambiguation by German upper case spelling can be done.



The metrics of precision and recall can not be applied. For accuracy the Lexicon Failures are dropped (as described in section 5.1), resulting into the sum of 119 words. Accuracy can be calculated as follows:

$$accuracy = \frac{correct}{correct + wrong} = 93.28\%$$

For the sake of completeness: If Lexicon Failures are counted as errors as well, accuracy computes to 74,0%.

7 Some Problems in Detail

7.1 Dictionary Coverage

The limited coverage of GermaNet as dictionary can also lead to wrong splits. The following example shows that what kept *BananaSplit* from working properly was the missing of an entry for *Klinke* in the lexicon:

Türklinke

[.N [.N Türke] [.B -e+∅] [.N Link] [.U -∅+∅] [.I -∅+e]]

7.2 Limiting POS

Performance could be improved by limiting the inflection handling procedure to work on the same parts-of-speech only: If the POS of a word changes during “deflection” what actually happens is derivation resp. “de-derivation”. The following example demonstrates how the word *tauben/A* (adjective, plural) is treated:

tauben

[.N [.N Taube] [.U -∅+∅] [.I -∅+n]]

The solution to this problem requires the input to be reliably POS-tagged which is not the case in the scenario of *BananaRelation*.

8 Conclusions

Measured with a random sample of 150 words from the TIGER corpus, the *BananaSplit* compound splitter achieves an accuracy of over 93%. Yet this is only valid under the condition that processing is restricted to nouns, verbs, and adjectives and that errors caused by missing entries in the dictionary (GermaNet) are excluded from the measurement procedure.

Köhn and Knight (2003) start off with a raw baseline of 94.2% accuracy. With their most sophisticated splitting method they achieve a remarkably high accuracy of 99.1%. The large difference between the accuracies of their raw baseline and the raw baseline in my experiment (34.45%) supposes that inflection must have been already treated before compound splitting.² *BananaSplit* operates on purely raw input and therefore performs less accurately.

Furthermore, poor GermaNet coverage and the missing handling of POS during inflection handling drag down accuracy as well. These points could be improved in the future.

References

- Brants, Sabine, Dipper, Stefanie, Hansen, Silvia, Lezius, Wolfgang, and Smith, George (2002). *The TIGER Treebank*. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol.
URL <http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/treeling2002.pdf>
- Köhn, Phillipp and Knight, Kevin (2003). *Empirical Methods for Compound Splitting*. In *EACL 2003*.
URL <http://people.csail.mit.edu/people/koehn/publications/compound2003.pdf>
- Langer, Stefan (1998). *Zur Morphologie und Semantik von Nominalkomposita*. In *Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache, KONVENS*, pp. 83–97.
URL <http://citeseer.ist.psu.edu/496465.html>

²They do not comment on that in their paper.