Simple Compound Splitting for German

Marion Weller-Di Marco

University of Stuttgart · Ludwig-Maximilians-University of Munich
dimarco@ims.uni-stuttgart.de

Introduction

- Compound: concatention of two or more words

  Apfel|baum (apple tree)
  Apfel|kuchen|rezept|sammlung (apple cake recipe collection)

- Productive word formation process
  → infinite amount of possible compounds

- Compound splitting useful for many NLP applications
  - Statistical Machine translation: translation of new compounds, better lexical coverage
  - Information retrieval: better generalization

- Splitting not trivial *Staubecken* → *Stau|becken* or *Staub|ecken*?

- Morphological operations on the modifier
  - *Bilder|Buch* → *Bild|Buch* transitional element
  - *Bücher|Regal* → *Buch|Regal* "Umlautung"

- Linguistically informed compound splitting with minimal resources
  - Morphological operations: learned from lemmatized data by mapping inflected forms to lemmas
  - Small set of hand-crafted rules for transitional elements
  - POS information for a flat analysis

    häuserfassade → haus_NN fassade_NN *house front*
    abfüllanlage → abfüllen_V anlage_NN *filling facility*

Modeling Transitional Elements

- Many modifier forms are part of the **inflectional inventory**: mostly plural or genitive forms

- Transitional elements/morphological operations for noun compounds Duden

Noun+Noun

add -en	Tat**en**drang	Tat|Drang	pl
add -n	Hase**n**braten	Hase|Braten	pl
add -ens	Herz**ens**güte	Herz|Güte	gen
add -ns	Frieden**s**vertrag	Frieden|Vertrag	gen
add -es	Kind**es**wohl	Kind|Wohl	gen
add -er	Büch**er**regal	Buch|Regal	pl
add -e	Hund**e**hütte	Hund|Hütte	pl
add -s	Museum**s**leiter	Museum|Leiter	gen
	Ansicht**s**karte	Ansicht|Karte	∅
rem. -e	Kirchturm	Kirch**e**|Turm	∅

Verb+Noun

add -en	Schreibmaschine schreib**en**|Maschine
add -n	Wanderweg wander**n**|Weg
context of *n, m*	
rem -e-	Rech**e**ngerät
add -en	rechn**en**|Gerät

Other+Noun

no modifications

Implemented Rules

Map inflected forms to lemma
→ approximate morphological operations

Bücher|regal
$Bücher_{Plural} \rightarrow Buch_{Lemma}$

Explicit modeling of transitional elements not contained in the inflectional inventory:

- Noun: *remove -s*
- Noun: *add -e*
- Noun: *remove -s*, *add -e*
- Verb: *add -en* (including deletion of *-e* in the context of *n,m*)
- Verb: *add -n*

No further rules needed

Splitting Method

- Frequency-based approach (Koehn&Knight 2003) extended with form-to-lemma mapping to handle compounding morphology

- Training data: frequency lists of tagged and lemmatized data
  - map inflected forms to lemmas
  - part-of-speech information to restrict splitting possibilities
    Gründer ('founder') ↛ *grün|der* ('green|the')

- **Splitting process**

input	breitflügelfledermaus_NN
split-1	breitflügel_XX fledermaus_NN
split-2	breit_ADJ flügel_NN fledermaus_NN
	'broad' *'wing'* *'bat'*

- Splitting possibilities are scored by the geometric mean of lemma frequencies

Categories

Modifier tags are restricted to

- **ADV** *wieder|aufforstung* 're|forestation'
- **ADJ** *alt|bestand* 'old|stock'
- **PART** *auf|preis* 'sur|charge'
- **V** *wandern|weg* 'hiking track'
- **NN** *apfel|kuchen* 'apple cake')
- **NE** *adam|apfel* 'adam's apple'

Additional "other" to add new category, for example neoclassical modifiers (e.g. *hydro-*)

Evaluation

	correct split	wrong split	not split	P	R	F
SMOR Split	45,054	2,914	3,262	93.93	87.94	90.84
Simple Split	46,905	4,012	313	92.12	91.56	91.84

- Test set: 51,230 binary split noun compounds

- SMOR-Split: contrastive splitter using the morphological resource SMOR Fritzinger et al. (2010)

- SMOR-Split
  precision-oriented conservative splitting, no splitting of lexicalized compounds or particles

- Simple Split
  recall-oriented splitting, covers most proper nouns

  Beaufort|Skala
  Bennett|Känguru

Conclusion

- Simple approach to compound splitting with minimal resources
- Morphological operations: approximated by form-to-lemma mapping
- Small set of rules for uncovered transitional elements
- Competitive results with a splitter relying on a high-quality morphological resource