Modelling sublexical analysis as sequence classification

Kay-Michael Würzner

wuerzner@bbaw.de

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Introduction

Words are not simply strings of characters. They bear an **internal structure** on multiple (partially hierarchically ordered) levels:

Morphology: Refers to the grammatical structure of words. Elements of the structural description are called *morphemes*.

Syllabification: Refers to the phonological structure of words. Elements of the structural description are called *syllables*.

Phonology: Refers to the phonological structure of syllables. Elements of the structural description are called *phonemes*.

Automatic analyses of particular words with respect to each of these levels are classical *natural language processing* (NLP) tasks. Existing approaches can be roughly divided into two types:

1. systems using **manually constructed rules** and
2. systems based on some **statistical model** automatically induced from training data.

With the increasing availability of manually annotated data (e.g. in professional lexical databases such as *CELEX* or community-driven projects such as *Wiktionary*), the latter have become the main focus of NLP research. However, rule-based systems still outperform statistical approaches in terms of correctness.

Inter-level dependencies

The different levels of sublexical representation are not independent of each other:

Word	*verifizieren*	*verirren*
Morphology[1]	ver ~ ifizier ~ en	ver + irr ~ en
Syllabification[2]	ve–ri–fi–zie–ren	ver–ir–ren
Phonology[3]	ˌvɛʁifiˈtsiːʁən	fɛɐ̯ˈʔɪʁən

1) ~ denotes a following suffix, + denotes a preceding prefix.
2) – denotes a syllable boundary.
3) The phonological representation is denoted using the International Phonetic Alphabet.

This German example illustrates, e.g., **the dependency of the syllable structure on the morphological structure**: German syllables are usually distributed following the *maximum onset principle*, effectively assigning the first r in *verifizieren* to the second syllable. However, this principle is violated in *verirren* due to the *stronger* rule that each prefix boundary co-occurs with a syllable boundary.

Comparing the phonological representations, shows that **the syllable structure in turn influences the pronunciations**: For example, the initial glottal stop in the second syllable of *verirren* is a consequence of the missing (overt) onset of that syllable.

Expectations

Both [1] and [2] use *conditional random fields* as the underlying type of statistical model. Their results are promising but far from optimal.

Expectation 1: Acquire the necessary knowledge to implement sequence classification for sublexical analysis with the methods of deep learning.

By now, there is no (statistical) approach which respects the dependencies between the different levels of sublexical word structuring.

Expectation 2: Investigate whether recurring neural networks might be an option to model these dependencies.

Sequence classification

Sublexical analysis may be treated as an instance of the **sequence classification problem**. I.e., given

- a set of symbols O and
- a set of classes C,

each symbol $o_i \in O$ in an observation string $o = o_1 \dots o_n$ **is mapped onto a class** $c_i \in C$ by determining the most probable string of classes $c = c_1 \dots c_n$ associated with o by an underlying stochastic model.

Morphological analysis

For the task of morphological analysis, O is defined to be the surface character alphabet itself. Following [1], the set of target classes is a "type-sensitive classification scheme" ($C = \{+, \#, \sim, 0\}$, where '+' indicates that a prefix morpheme ends at the current position, '#' indicates that a free morpheme starts with the following position, '$\sim$' indicates that a suffix morpheme starts with the following position, and 0 indicates that there is no morpheme boundary after the current position).

G e f o l g s l e u t e n
o + o o o ~ # o o o o ~ o

Syllabification

For the task of syllabification, O is again defined to be the surface character alphabet itself. The set of target classes is binary ($C = \{0, 1\}$), leading to a classifier which predicts for every position i whether or not there is a syllable boundary following (rsp. preceding) the observed symbol at position i of the input word.

G e f o l g s l e u t e n
o 1 o o o o 1 o o 1 o o o

Grapheme-Phoneme conversion

In contrast to the aforementioned tasks, the grapheme-phoneme correspondence can not be (straightforwardly) modelled as a $1 : n$ mapping. In [2], a constrained-based alignment is proposed to deal with that problem: a grapheme alphabet Σ_G, a phoneme alphabet Σ_P, and a finite set $M \subset (\Sigma_G^+ \times \Sigma_P^+)$ relating grapheme substrings and their potential phonemic realizations are used to generate o and c.

$$M = \left\{ \begin{matrix} \text{p} : /\text{p}/,\ \text{h} : /\text{h}/,\ \text{ph} : /\text{f}/,\ \text{ö} : /\text{ø:}/,\ \text{ö} : /\text{œ}/ \\ \text{n} : /\text{n}/,\ \text{i} : /\text{ɪ}/,\ \text{k} : /\text{k}/,\ \text{s} : /\text{s}/,\ \text{x} : /\text{ks}/ \end{matrix} \right\}$$

Grapheme segmentations = { p—h—ö—n—i—x—, ph—ö—n—i—x— }

Phoneme segmentations = { f_ø:_n_ɪ_ks_, f_ø:_n_ɪ_k_s_ }
Alignment = { ph—ö—n—i—x— : f_ø:_n_ɪ_ks_ }

References

[1] K.-M. Würzner and B. Jurish. Dsolve – morphological segmentation for german using conditional random fields. In *Systems and Frameworks for Computational Morphology*, volume 537 of *Communications in Computer and Information Science*, pages 94–103. Springer, 2015.

[2] K.-M. Würzner and B. Jurish. A hybrid approach to grapheme-phoneme conversion. In *Proceedings of the 12th International Workshop on Finite State Methods and Natural Language Processing*, 2015.