CaMEL: Case Marker Extraction without Labels

Leonie Weissweiler, Valentin Hofmann, Masoud Jalili Sabet, Hinrich Schütze
Deep Cases

- Case marks the role of a Noun Phrase (NP) in a given sentence
- Deep Cases (Filmore, 1968) are language-universal and more fine grained

| Deep Case   | Description                                         | Example                             |
|-------------|------------------------------------------------------|-------------------------------------|
| Nominative  | The subject of the sentence                         | He is the Messiah!                 |
| Genitive    | An entity that possesses another entity             | Are you the Judean People’s Front? |
| Recipient   | A sentient destination                              | I gave the gourd to Brian.         |
| Accusative  | The direct object of the sentence                   | Consider the lilies.               |
| Locative    | The spatial or temporal position of an entity       | They haggle in the market.         |
| Instrumental| The means by which an activity is carried out       | The graffiti was written by hand.  |
Overlapping Case Systems in Parallel Text

Case markers, case systems and deep cases are not mapped one-to-one:

- Case polysemy: one case, several deep cases
- Case homonymy: several cases, one marker
- Case synonymy: one case, several markers

→ **Key idea:** we can gain information about the deep case of an NP involving –ibus in a given context by looking at the case markers in its Russian translation.
Contributions

- We introduce **CaMEL**: **Ca**se **M**arker **E**xtraction without **L**abels 🐫, the task of extracting the case markers for unannotated parallel text
- We propose a simple method that is efficient, doesn’t require training, and generalises well to new languages
- We automatically construct a silver standard based on UniMorph data and evaluate our method, achieving 45% average F1 over 19 languages
- We demonstrate two first ways of using the extracted case markers
Our Method

Parallel Corpus with alignments

NP Annotation

NP Projection

Candidate Set Creation

Frequency Filtering

Inside/Outside Filtering

Restriction to Word Endings
NP Annotation and Projection

I am the fine sheperd; the fine sheperd surrenders his soul on behalf of the sheep.
Ich bin der vortreffliche Hirte; der vortreffliche Hirte gibt seine Seele zugunsten der Schafe hin.

I am the good sheperd; the good sheperd sacrifices his life for the sheep.
Ich bin der vortreffliche Hirte; der vortreffliche Hirte gibt seine Seele zugunsten der Schafe hin.
Candidate Set Creation

- We now have a frequency list of words inside of NPs and outside of NPs for each language.
- We move words with a higher relative frequency inside of NP to $I_l$ and all others to $O_l$.
- From $I_l$, we generate our candidate set, with all character n-grams from all words in $I_l$, e.g. *ovibus* ‘sheep’ $\Rightarrow$ $ovi$, $ibus$, but also $ovibus$, and i etc.
Filtering of the Candidate Set

- Frequency Filtering: we filter out all candidates with a frequency lower than a threshold
- Inside/Outside Filtering
  - we conduct a Fisher’s Exact Test on the frequencies of a candidate inside and outside of NPs
  - Question: does this candidate occur more frequently inside than outside of NPs?
  - use the resulting p-value and odds ratio for filtering
- Restriction to word endings
Silver Standard

- Automatically created from paradigms in UniMorph
- Covers 19 languages
- Emphasis on precision rather than recall

| Nominative Singular | inflected forms | unused information |
|---------------------|-----------------|--------------------|
| base                | suffix          |                    |
| Abfl ug             | es              | N NOM SG           |
| Abfl ug             | N GEN SG        |
| Abfl ug             | N DAT SG        |
| Abfl üge            | N ACC SG        |
| Abfl ügen           | N NOM PL        |
| Abfl üge            | N GEN PL        |
| Abfl üge            | N DAT PL        |
| Abfl üge            | N ACC PL        |

Abflug
Quantitative Evaluation

We achieve 54% average precision, 41% average recall and 45% average F1 over all 19 languages

| Intersection | Algorithm Only | Silver Standard Only |
|--------------|---------------|----------------------|
| u, я, ом, ого, о, в, ой, и, ми, ам, ей, ю, й, ов, ых, а, м, х, ами | ий, ные, ое, ение, ии, го, ый, ка, ые, к, ки, ия, ние, й, ния, ие | ыми, ах, ев, ыям, ому, йа, н, ыях, ями, ям, е, ях, ьев, ем, ым, ья-ми |
| i, ja, om, ogo, o, v, oj, i, mi, am, ej, ju, y, ov, yx, a, m, x, ami | ij, nye, oe, enie, ii, go, yj, ka, ye, k, ki, ija, nie, j, nija, ie | ymi, ax, ev, 'jam, omu, 'ja, n, 'jax, jam, jami, jam, e, jax, 'ev, em, um, 'jami |
Manual Qualitative Evaluation

- **domibus** – дворцах/dvorcax – **Location**
  → ‘in the houses’
- **operibus bonis** – добрыми делами/dobrymi delami – **Instrumental**
  → ‘through the good deeds’
- **patribus** – предкам/predkam – **Recipient**
  → ‘for/to the parents’
Semi-Automated Qualitative Evaluation

• Generate NP-word co-occurrence matrix over the NP vocabulary of all languages
• Reduce with t-SNE
• Here: NPs with Latin –ibus, coloured by occurrence of Polish ach$ (LOC) and –om$ (DAT)
• → we can cluster NPs semantically by their deep case
Conclusion

We have

• introduced the new task of **Case Marker Extraction without Labels CaMEL**
• compiled an automatically created silver standard for this task covering 19 languages
• presented a simple and efficient method leveraging alignments and achieving 45% average F1
• demonstrated two ways in which the retrieved case markers can be used to investigate deep case
Thank you for listening!