IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Miriam Amin, Peter Fankhauser, Marc Kupietz, Roman Schneider

DATA-DRIVEN IDENTIFICATION OF IDIOMS IN SONG LYRICS

Leibniz Gemeinschaft

INTRODUCTION

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Problem statement

- Automatic identification of idioms beneficial for information extraction, retrieval, summarization and translation
- Idioms as 'pain in the neck for NLP' (Sag et al. 2002)

Research objective

- Cover idiom characteristics with an innovative set of quantitative features
- Apply and evaluate machine-learning classifiers for an idiomatically rich corpus

HYPOTHESIS

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Idioms are characterised by:

1. High degree of formal fixedness/phraseness
2. Unusual usage
3. Unusual context

DATASET

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Corpus of German Song Lyrics (Schneider 2020)

- Approx. 1.800.000 tokens in 5.000 songs
- Random selection of approx. 10.000 2-6-grams from the corpus
- Manual annotation (idiom/non-idiom) and removal of unclear cases
- Final dataset: 542 idioms and 8.697 non-idioms

SYNTAGMATIC FEATURES

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- **Formal Fixedness:**
  - SY_C1: Count-based collocation measures calculated on German reference corpus (DeReKo, Kupietz et al. 2010)
  - SY_C2: Count-based collocation measures calculated on Corpus of German Song Lyrics
- **Unusual usage:**
  - SY_W: Predictive collocation measures calculated with word2vec
- **Approach:**
  - SY_R: Rank-based collocation measures
  - All measures are calculated taking the average over all pairs of words w_i , w_2 in an idiom candidate of size $/w/$

Holy cow vs. *cow faces*

Holy cow vs. *grazing cow*

CONTEXT AND OTHER FEATURES

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Unusual context

- CO: Measure semantic similarity between idiom candidate and context
- Calculated as mean cosine similarity between all words w_i in the idiom candidate of size /w/ and words c_j in the left/right context of size /c/

Pearls before swine

Other features

- O: Various word counts

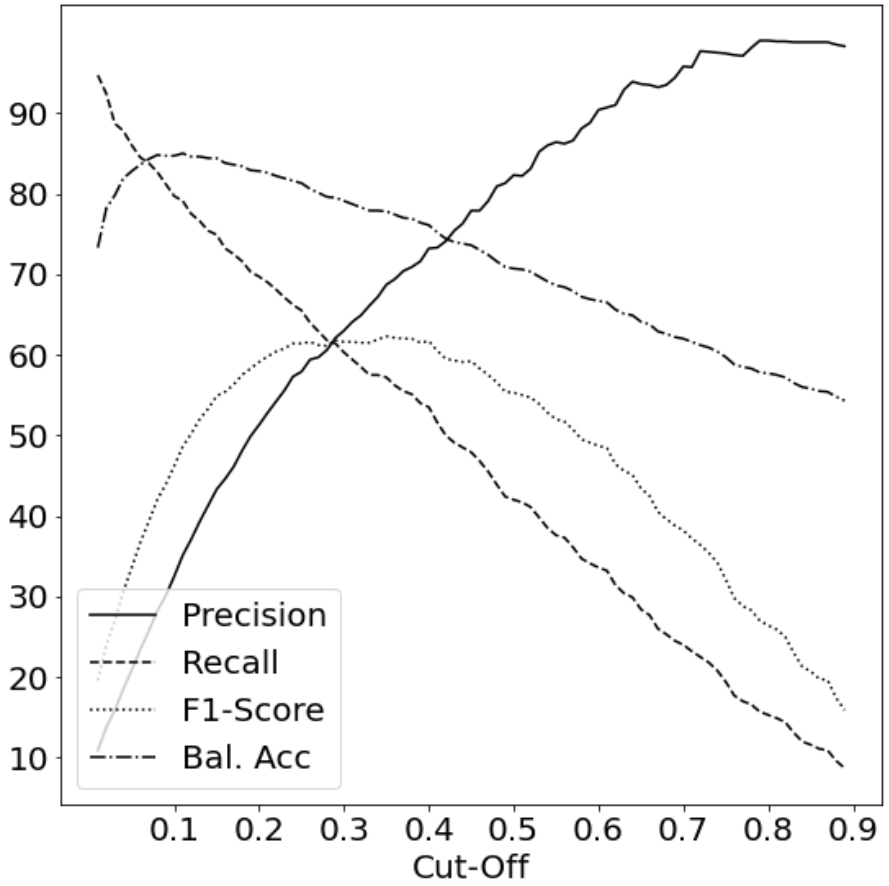
METHODS

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Random Forest Classifier
- 5-fold cross validation

RESULTS 1/2

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE



Trade-off curves for Random Forest cut-off

RESULTS 2/2

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Feature set	Precision	Recall	F1-Score	Bal. Acc.
All features	62.7	59.9	61.3	78.9
SY_C1	44.2	38.7	41.2	67.8
SY_C2	32.9	30.6	31.7	63.4
SY_W	39.2	24.9	30.3	61.3
SY_R	31.2	28.0	29.5	62.1
CO	11.8	7.4	9.1	52.0
O	0.0	0.0	0.0	50.0
w/o SY_C1_R	55.8	48.9	52.1	73.2
w/o SY_C2	60.3	53.3	56.5	75.6
w/o SY_W_R	61.0	58.7	59.8	78.2
w/o SY_R	63.0	60.9	61.9	79.3
w/o CO	59.9	60.3	60.1	78.9
w/o O	61.0	55.9	58.3	76.8

SY_C	Count-based collocation measures	**Formal fixedness**
SY_W	Predictive collocation measures	**Unusual usage**
SY_R	Rank-based collocation measures	
CO	Context similarity	**Unusual context**
O	Other	

Performance of different feature sets in a Random Forest with cutoff=0.3.

CONCLUSIONS

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Count-based collocation measures indeed characterize idioms' **formal fixedness**
- Predictive collocation measures model **unusual usage**
- Context features are able to model **unusual context**

DISCUSSION

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Strengths

- Features do not require intensive preprocessing
- Works with minimal context
- Detects even idioms that were overlooked by manual annotation

Hinter Gitterstäben
(lit. 'behind thick bars')
In prison

Limitations

- Method does not work for idioms that consist of only one context word (plus stopwords)
- Works worse for novel idioms
- Does not work when the idiomatic use is the dominant use

Das A und O
(lit. 'A and O / alpha and omega')
The most important part

FUTURE WORK

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Apply approach on bigger dataset
- Experiment with additional features

REFERENCES

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Marc Kupietz, Cyril Belica, Holger Keibel, and An- dreas Witt. 2010. The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In *Proceedings of the Seventh International Conference On Language Resources And Eval- uation (LREC'10)*, page 1848–1854, Valletta / Paris. European Language Resources Association (ELRA).
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In Computational Linguistics and Intelligent Text Processing, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Roman Schneider. 2020. A corpus linguistic perspective on contemporary german pop lyrics with the multi-layer annotated "songkorpus". In Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11- 16, 2020, pages 842–848. European Language Resources Association.