POS and Morphological Tagging & Lemmatizing

Luisa Berlanda & Alexander Fraser

CIS, Ludwig-Maximilians-Universität München

Computational Morphology and Electronic Dictionaries
SoSe 2016
2016-06-27

Outline

1. Tagging

2. Lemmatizing

3. POS-Tagging for spoken language corpora

Outline

1. Tagging

2. Lemmatizing

3. POS-Tagging for spoken language corpora

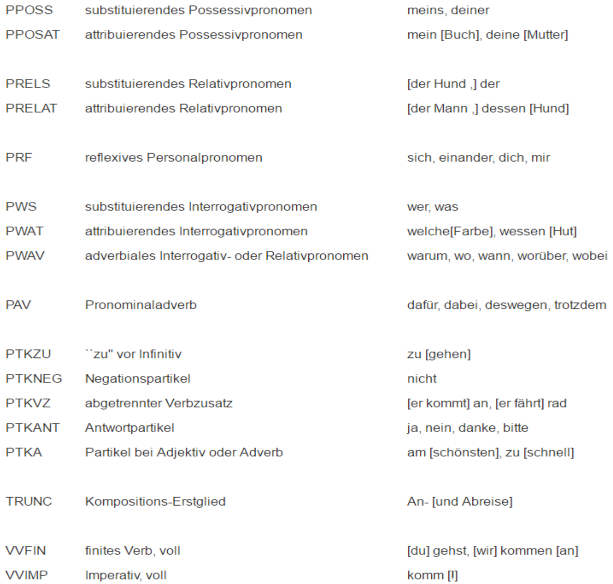
Basic principles

What is POS tagging?

- each word has a word class categorie
  - e.g. noun, verb, adjective, adverb...
- identification and assignation of word class categories
  - given a word form within a corpus
  - often using lexical and contextual information
- difficulty of tagging a corpus with POS tags
  - ⇒ single tokens are often **ambigous**

POS Tags

some Tags of the Stuttgart-Tübingen-Tagset (STTS)

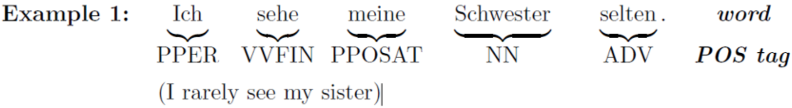


PPOSS	substituierendes Possessivpronomen	meins, deiner
PPOSAT	attribuierendes Possessivpronomen	mein [Buch], deine [Mutter]
PRELS	substituierendes Relativpronomen	[der Hund ,] der
PRELAT	attribuierendes Relativpronomen	[der Mann ,] dessen [Hund]
PRF	reflexives Personalpronomen	sich, einander, dich, mir
PWS	substituierendes Interrogativpronomen	wer, was
PWAT	attribuierendes Interrogativpronomen	welche[Farbe], wessen [Hut]
PWAV	adverbiales Interrogativ- oder Relativpronomen	warum, wo, wann, worüber, wobei
PAV	Pronominaladverb	dafür, dabei, deswegen, trotzdem
PTKZU	``zu" vor Infinitiv	zu [gehen]
PTKNEG	Negationspartikel	nicht
PTKVZ	abgetrennter Verbzusatz	[er kommt] an, [er fährt] rad
PTKANT	Antwortpartikel	ja, nein, danke, bitte
PTKA	Partikel bei Adjektiv oder Adverb	am [schönsten], zu [schnell]
TRUNC	Kompositions-Erstglied	An- [und Abreise]
VVFIN	finites Verb, voll	[du] gehst, [wir] kommen [an]
VVIMP	Imperativ, voll	komm [!]

Basic principles

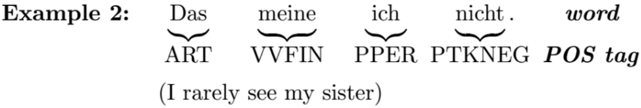
Example POS tag

- "meine" could be
  - ⇒ VVFIN or
  - ⇒ PPOSAT



Example 1: $\underbrace{\text{Ich}}_{\text{PPER}}$ $\underbrace{\text{sehe}}_{\text{VVFIN}}$ $\underbrace{\text{meine}}_{\text{PPOSAT}}$ $\underbrace{\text{Schwester}}_{\text{NN}}$ $\underbrace{\text{selten}}_{\text{ADV}}$. $\underbrace{\textbf{\textit{word}}}_{\textbf{\textit{POS tag}}}$

(I rarely see my sister)|



Example 2: $\underbrace{\text{Das}}_{\text{ART}}$ $\underbrace{\text{meine}}_{\text{VVFIN}}$ $\underbrace{\text{ich}}_{\text{PPER}}$ $\underbrace{\text{nicht}}_{\text{PTKNEG}}$. $\underbrace{\textbf{\textit{word}}}_{\textbf{\textit{POS tag}}}$

(I rarely see my sister)

- the context of a word in a sentence is crucial!

Basic principles

What is morphological tagging?

- assigning additional morphological information to each token
  - e.g gender, case, person, tense..
- very important for morphologically rich languages
- one POS tag can have different morphological analyses
  - Da gehen sie alle entlang.

  ⇒ 3.person, pl, Nom.
  - Da geht sie immer entlang.

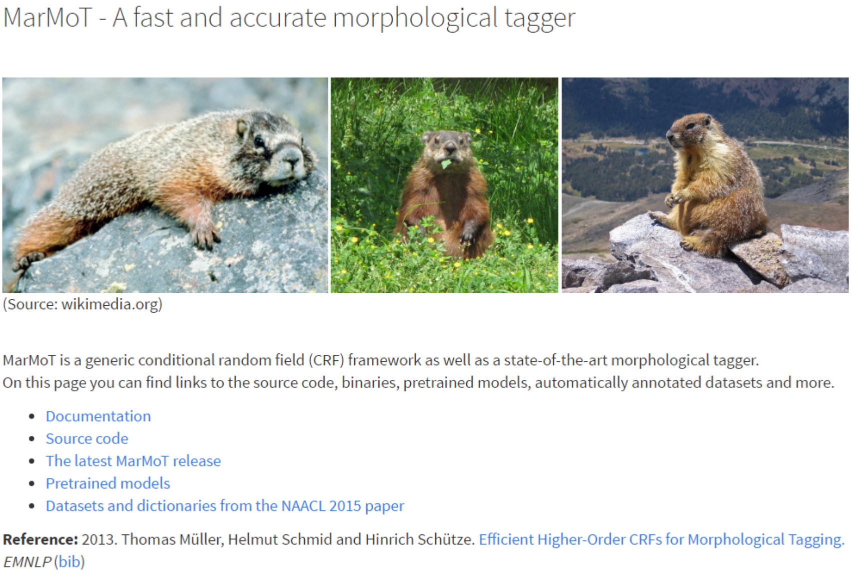
  ⇒ 3.person, sg, Nom.

Tools

POS and morphological Tagger

- MarMoT
- Conditional-Random-Field tagger, developed by Müller, Schmid and Schütze in 2013
- uses pruning, stochastic gradient descent training
  ⇒ applicable for huge tagsets
- available at http://cistern.cis.lmu.de/marmot/

MarMoT



MarMoT - A fast and accurate morphological tagger

(Source: wikimedia.org)

MarMoT is a generic conditional random field (CRF) framework as well as a state-of-the-art morphological tagger.
On this page you can find links to the source code, binaries, pretrained models, automatically annotated datasets and more.

- Documentation
- Source code
- The latest MarMoT release
- Pretrained models
- Datasets and dictionaries from the NAACL 2015 paper

Reference: 2013. Thomas Müller, Helmut Schmid and Hinrich Schütze. Efficient Higher-Order CRFs for Morphological Tagging. *EMNLP* (bib)

Outline

1. Tagging

2. Lemmatizing

3. POS-Tagging for spoken language corpora

Basic principles

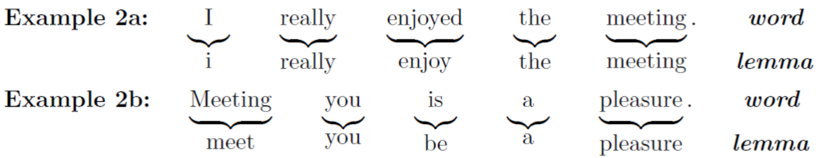
What is lemmatizing?

- lemma = uninflected word form
- each word has a lemma
- lemmatizing is similar to stemming
- groups word forms of the same inflectional paradigm together and then assigns the lemma

Basic principles

Example lemma

- "meeting" could be
  - ⇒ meet or
  - ⇒ meeting



Example 2a:	$\underbrace{\text{I}}_{\text{i}}$	$\underbrace{\text{really}}_{\text{really}}$	$\underbrace{\text{enjoyed}}_{\text{enjoy}}$	$\underbrace{\text{the}}_{\text{the}}$	$\underbrace{\text{meeting}}_{\text{meeting}}$.	***word*** ***lemma***
Example 2b:	$\underbrace{\text{Meeting}}_{\text{meet}}$	$\underbrace{\text{you}}_{\text{you}}$	$\underbrace{\text{is}}_{\text{be}}$	$\underbrace{\text{a}}_{\text{a}}$	$\underbrace{\text{pleasure}}_{\text{pleasure}}$.	***word*** ***lemma***

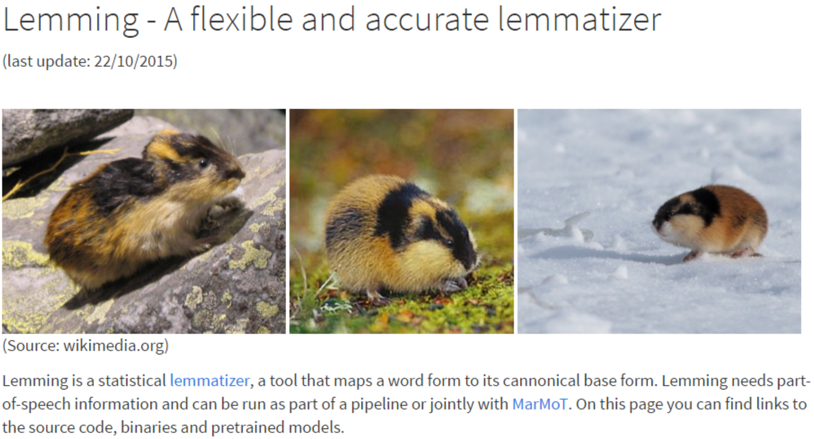
- the context is crucial again

Tools

Lemmatizer

- LEMMING
- state of the art token-based lemmatizer, developed by Müller et al., 2015
- statistical, token based approach
- modular log-linear model
- needs an annotated corpus with gold standard tags as a prerequisite
- use of arbitrary global features enables lemmatizing of unknown words
- available at http://cistern.cis.lmu.de/lemming/

LEMMING



Lemming - A flexible and accurate lemmatizer

(last update: 22/10/2015)

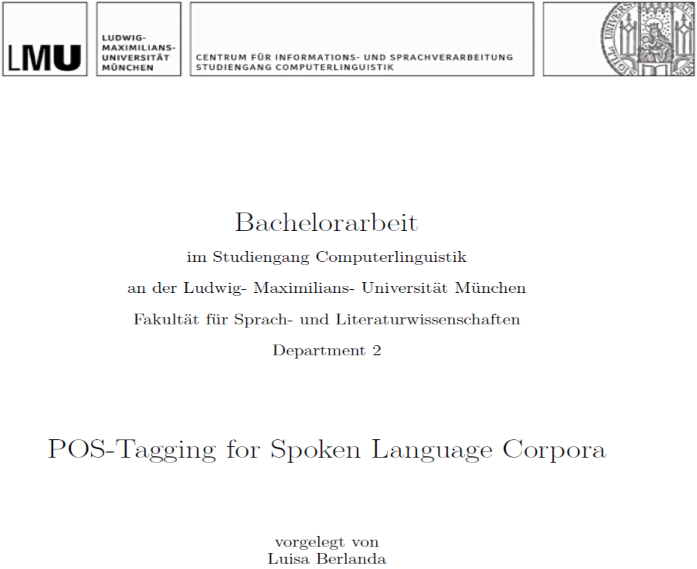
(Source: wikimedia.org)

Lemming is a statistical lemmatizer, a tool that maps a word form to its cannonical base form. Lemming needs part-of-speech information and can be run as part of a pipeline or jointly with MarMoT. On this page you can find links to the source code, binaries and pretrained models.

Outline

1. Tagging
2. Lemmatizing
3. POS-Tagging for spoken language corpora

Bachelor Thesis



LMU

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Bachelorarbeit

im Studiengang Computerlinguistik

an der Ludwig- Maximilians- Universität München

Fakultät für Sprach- und Literaturwissenschaften

Department 2

POS-Tagging for Spoken Language Corpora

vorgelegt von
Luisa Berlanda

x

POS-Tagging for Spoken Language Corpora

What is done in this work?

- Using a POS tagger and lemmatizer designed for written language for the analysis of a spoken language corpus
- Conducting several experiments with the spoken data
- Performing an error analysis

- POS tagger: MarMoT (Müller, Schmid and Schütze, 2013)
- Lemmatizer: LEMMING (Müller et al., 2015)
- Corpus: FOLK-Gold (Westpfahl and Schmidt, 2016)

Motivation

Why is this an interesting topic?

- POS tagging and lemmatizing are basic tasks for NLP
  ⇒ Prerequisite for many applications
- CL is very active in the world wide web
  ⇒ Spoken language as new standard medium of communication
- Adapting existing taggers has a great benefit
- The new gold standard corpus FOLK-Gold enables the training on spoken data

Spoken language phenomena

Differences to written language

- Missing bondary information
- Disfluencies
  - ⇒ Discourse markers
  - ⇒ Interjections
  - ⇒ Speech repairs
  - ⇒ Silent / filled pauses
- What can a spoken language corpus contain?
  - ⇒ Audio-files, raw transcript, phonetic transcript, normalized form, metadata

FOLK-Gold

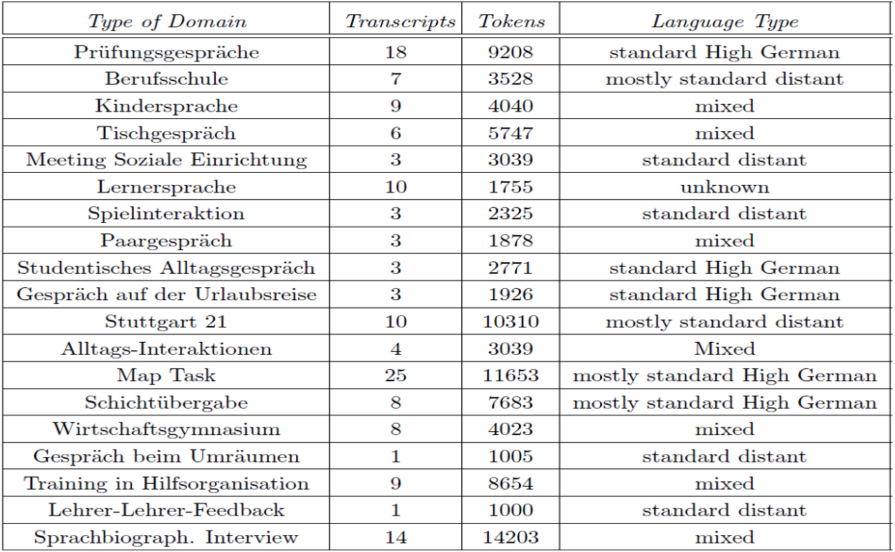
How does the corpus look like?

- FOLK-Gold Corpus (Westpfahl and Schmidt, 2016)
- Annotated gold standard corpus for German
- ca. 100.000 tokens
- annotation layers

  ⇒ transcription, normalization, lemma and POS tag of each token
- Utterances split by pauses $>$ 0.2 sec.
- 19 different domains

FOLK-Gold

Overview of the domains



Type of Domain	*Transcripts*	*Tokens*	*Language Type*
Prüfungsgespräche	18	9208	standard High German
Berufsschule	7	3528	mostly standard distant
Kindersprache	9	4040	mixed
Tischgespräch	6	5747	mixed
Meeting Soziale Einrichtung	3	3039	standard distant
Lernersprache	10	1755	unknown
Spielinteraktion	3	2325	standard distant
Paargespräch	3	1878	mixed
Studentisches Alltagsgespräch	3	2771	standard High German
Gespräch auf der Urlaubsreise	3	1926	standard High German
Stuttgart 21	10	10310	mostly standard distant
Alltags-Interaktionen	4	3039	Mixed
Map Task	25	11653	mostly standard High German
Schichtübergabe	8	7683	mostly standard High German
Wirtschaftsgymnasium	8	4023	mixed
Gespräch beim Umräumen	1	1005	standard distant
Training in Hilfsorganisation	9	8654	mixed
Lehrer-Lehrer-Feedback	1	1000	standard distant
Sprachbiograph. Interview	14	14203	mixed

Table 1: Overview of the different domains included in FOLK-Gold

FOLK-Gold

Dimensions in the corpus

- Language type
  - ⇒ 41.6% non-standard speech (regional variants, vernaculars)
  - ⇒ 46.7% standard language
  - ⇒ 11.8% mixed
- Conversation type
  - ⇒ 54.2% formal conversations
  - ⇒ 45.8% informal conversations
- Level of interaction
  - ⇒ 59.6% disciplined conversation
  - ⇒ 40.4% interactive conversations
- 3% child language
- 1.3% speech of non-native speakers

Experiments

Baseline

- MarMoT and LEMMING in pipeline model
- Data in .tsv file
- Tags for special phenomena: PAUS, BREA, NONPH, .
- Division in train, validation, test set: 70:15:15

- Training on TIGER data
  ⇒ + morphological dictionary, + MarLiN cluster file

- Testing on transcribed forms
- Testing on normalizes forms

Experiments with FOLK-Gold

- Trainig on FOLK-GOLD normalized forms
- Testing with
  - ⇒ Spoken forms
  - ⇒ Normalized forms
  - ⇒ Without inconsistencies
  - ⇒ With cluster and morphological file

Results

What can be concluded from the experiments

- Training on FOLK-Gold best setting
- Normalized forms better than spoken forms (testing)
- Morphological information is helpful

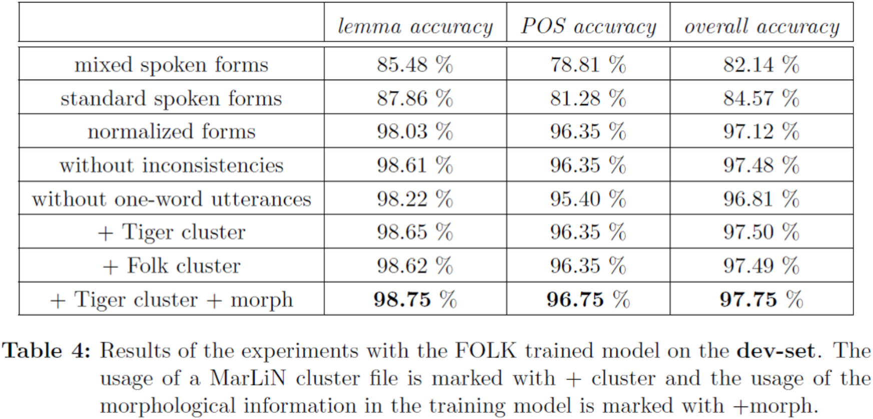
- Language type

  ⇒ standard spoken forms best
- Domains

  ⇒ no great impact on results, huge samples with standard spoken forms best

Results

Experiment results on the dev set



	lemma accuracy	*POS accuracy*	*overall accuracy*
mixed spoken forms	85.48 %	78.81 %	82.14 %
standard spoken forms	87.86 %	81.28 %	84.57 %
normalized forms	98.03 %	96.35 %	97.12 %
without inconsistencies	98.61 %	96.35 %	97.48 %
without one-word utterances	98.22 %	95.40 %	96.81 %
+ Tiger cluster	98.65 %	96.35 %	97.50 %
+ Folk cluster	98.62 %	96.35 %	97.49 %
+ Tiger cluster + morph	**98.75** %	**96.75** %	**97.75** %

Table 4: Results of the experiments with the FOLK trained model on the **dev-set**. The usage of a MarLiN cluster file is marked with + cluster and the usage of the morphological information in the training model is marked with +morph.

Outlook

Further Experiments

- Training on the spoken forms
- Adding a morphological dictionary tailored to spoken language
- Adding a spoken-form dictionary
- Adding a name lexicon
- Huge German dictionary as external ressource
- Annotating the corpus with fine-grained morphological tags
- Exploring other dimensions
  - ⇒ conversation type, level of interaction

Outlook

Above this work

- Use meta-information, e.g. time-stamps
- Other boundaries
- Overlapping speech and speech repairs
- Varie the domain of written language

References

- S. Brants, S. Dipper, P. Eisenberg, S. Hansen, E. König, W. Lezius, C. Rohrer, G. Smith, H. Uszkoreit. TIGER: Linguistic Interpretation of a German Corpus. In: Journal of Language and Computation, 2004:2, pp. 597-620.
- P.A. Heeman, J.F.Allen. Speech Repairs, Intonational Phrases, and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialogue. In: Computational Linguistics, 25:4, 1999, pp. 527-571.
- Institut für Deutsche Sprache. DGD. Datenbank für Gesprochenes Deutsch. April 1, 2016, http://dgd.ids-mannheim.de.
- T. Müller, R. Cotterell, A. Fraser, H. Schütze. Joint Lemmatization and Morphological Tagging with LEMMING.In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, September 2015, Lisbon, Portugal, Association for Computational Linguistics, 2015, pp. 2268-2274.
- T. Müller, H. Schütze. Robust Morphological Tagging with Word Representations. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies, May 31 -June 5, 2015, Denver, Colorado, Association for Computational Linguistics, 2015, pp. 526 -536.

References

- T. Müller, H. Schmid, H. Schütze. Efficient Higher-Order CRFs for Morphological Tagging. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), October 18–21, 2013, Seattle, USA, Association for Computational Linguistics (ACL), 2013, pp. 322-332.
- T. Schmidt. Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD. In: Gesprächsforschung -Online-Zeitschrift zur verbalen Interaktion, 15, 2014, pp. 196-233.
- T. Schmidt. EXMARaLDA and the FOLK tools -two toolsets for transcribing and annotating spoken language. In: Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12), May 21-27, 2012, Istanbul, Turkey, pp. 236 -240.
- T. Schmidt, K. Wörner. EXMARaLDA -Creating, Analyzing and Sharing Spoken Language Corpora for Pragmatics Research. In: Pragmatics -Quarterly Publication of the International Pragmatics Association, 19:4, 2009, pp. 565-582.
- S. Westpfahl, T. Schmidt. FOLK-Gold -A GOLD standard for Part-of-Speech-Tagging of Spoken German. In: Proceedings of the Tenth conference on International Language Resources and Evaluation (LREC'16), May 23 -28, Portoroz, Slovenia, ELRA, 2016, pp. 1493 -1499.
- S. Westpfahl, T. Schmidt. POS für(s) FOLK -Part of Speech Tagging des Forschungs-und Lehrkorpus Gesprochenes Deutsch. In : Journal for Language Technology and Computational Linguistics (JLCL), 28:1, 2013, pp. 139-153.

Thank you for your attention.