Computational Morphology
FOU17

Harald Hammarström
Uppsala University
harald.hammarstrom@lingfil.uu.se

30 Aug 2017 Uppsala

Computational Morphology

Break words into meaningful units, i.e., morphemes

flickornas

flick-or-na-s

antidisestablishmentarianism

anti-dis-establish-ment-arian-ism

Useful or even crucial for many downstream tasks in Information Retrieval, Machine Translation etc

Hand-crafted Rules

- Write hand-crafted rules that describe the legal stem+ending combinations

beg	Vinf	+V:0	#
fox	Vinf	+V+3P+Sg:^s	#
make	Vinf	+V+Past:^ed	#
panic	Vinf	+V+PastPart:^ed	#
watch	Vinf	+V+PresPart:^ing	#

- Typically rules are written in a finite-state formalism for efficient analysis and generation (once compiled) read Hulden (2009), for the library FOMA and the more practically oriented tutorial `http://foma.sourceforge.net/lrec2010/`.

Supervised Learning of Morphology

- Feed examples of [inflected, *features*] $\Rightarrow$ stem
- To a supervised ML algorithm: Support Vector Machine, Decision-Tree, k-NN, Neural Network etc
- As features one may supply the k first and last characters of the inflected form

`Input features:` flickornas, *f*, *fl*, *fli*, *flic*, *flick*, *ornas*, *rnas*, *nas*, *as*, *s*

$\Downarrow$

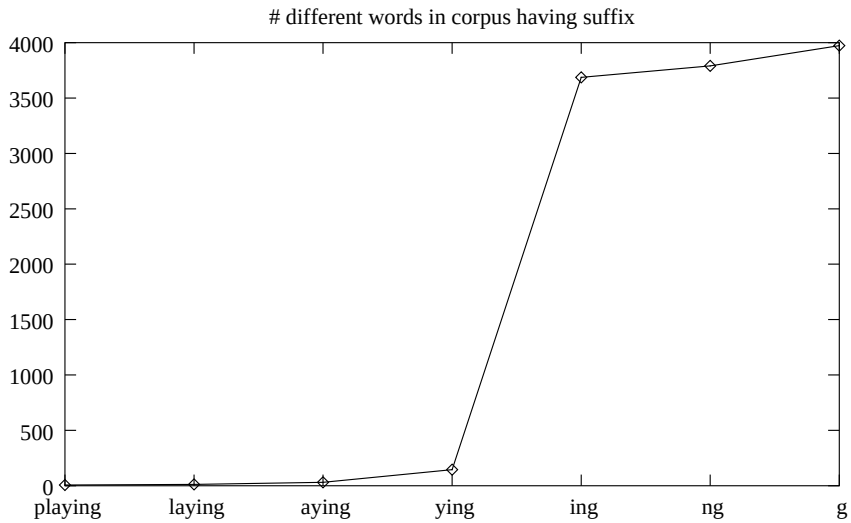
`Output:` flick

- Read Kann and Schütze (2016) and Chrupala (2008) chapter 6 and and see `https://sites.google.com/site/morfetteweb/`.

Unsupervised Learning of Morphology

- Input: Just raw unannotated text data in large amounts
- Output: Segmented text data

Why would this work at all?



- Frequency asymmetries may be exploited to extract affixes
- Frequency asymmetries which affixes occur on which stems and vice versa
- Read Moon et al. (2009) for a concrete system and Hammarström and Borin (2011) for an overview.

Morphology Learning with Parallel Text

English	and there was evening and there was morning on the third day
Swedish	och det vart afton och det vart morgon den tredje dagen
Maori	a ko te ahiahi ko te ata he ra tuatoru
West Greenlandic	Taava unnunngorpoq ullaanngorlunilu ullut pingajuat

- Can morphology learning be helped if you have parallel text, read Snyder and Barzilay (2008)?
- What if you have segmentation in one of the languages?
- What can you get out of parallel texts more generally with neural embeddings, read Östling and Tiedemann (2017)

Further Twists

- Concatenative versus non-concatenative morphology? Read, e.g., Khaliq (2015)

	Arabic			Finnish	
	3p.sg.per	3p.sg.impf		nom.sg.	gen.sg
'write'	*kataba*	*yaktubu*	'flower'	*kukka*	*kuka-n*
'kill'	*qatala*	*yaqtulu*	'girl'	*tyttö*	*tytö-n*

- Just do segmentation or infer inflectional paradigms, read Chan (2006)
- Compound splitting read, e.g., Ma et al. (2016)
- Include semantics (somewhere) in the morphology learning, read Deerwester et al. (1990) for Latent Semantic Indexing (LSI) or Mikolov et al. (2013) for Word2Vec

Training/Test Data and Libraries

- UniMorph: Various amounts of data for 51 (!) languages, see `https://unimorph.github.io/index.html`

  atxiki betxekie V;ARGABSSG;ARGIOPL;IMP
  atxiki betxekik V;ARGABSSG;ARGIOSG;ARGIOMASC;IMP
  atxiki betxekin V;ARGABSSG;ARGIOSG;ARGIOFEM;IMP
  ...

- Swedish: SALDO `https://spraakbanken.gu.se/swe/resurs/saldom`, English, German, Finnish, Turkish, see `http://morpho.aalto.fi/events/morphochallenge/`
- Parallel Bible texts (verse aligned) for appx 1000 languages, see `http://paralleltext.info/data/`
- FOMA (for hand-written rules) `https://fomafst.github.io`
- Word2Vec in Řehůřek and Sojka (2010)

Some Project Suggestions #1

Hand-crafted Morphological Analyzer: Compose rules manually to describe (a subset of) the morphology of a chosen language. If you use an existing framework you get a lot for free for a relatively small learning threshold. The research aspect is to device a concise set of rules and reuse/invent a framework that allows efficient generation and analysis.

Supervised Morphological Learner: Devise a supervised Machine Learning algorithm to learn the (re/un-)inflection for a chosen language/set of languages with a dataset of input-output pairs or a dataset constructed by yourself. The research aspect is to engineer an algorithm and choose a suitable representation and set of features.

Some Project Suggestions #2

Unsupervised Morphological Learner: Devise an unsupervised Machine Learning algorithm to learn the (re/un-)inflection for a chosen language/set of languages with a dataset of input-output pairs or a dataset constructed by yourself. The research aspect is to engineer an algorithm and choose a suitable representation and set of features.

Morphology Learning with Semantics: Most morphological learning systems are oblivious to semantics, i.e, they have no idea that *horse*/*horses* are semantically related but *stop*/*top* are not. Presumably they could improve with this knowledge. Devise a supervised/unsupervised morphological learner which makes use of semantics, for example, that which is obtained by distributional analysis in a corpus. The research aspect is how to integrate the information from semantics towards the target language morphological analysis.

Some Project Suggestions #3

Morphology Learning with Parallel Text: Devise a supervised/unsupervised morphological learner which makes use of parallel text. The research aspect is how to integrate the information from other language(s) and their links towards the target language morphological analysis.

Compound Splitting: Devise a supervised/unsupervised splitter for compounds, i.e., when two lexical items may be compounded and written together (e.g., raincoat). This is a slightly different problem compared to that of morphology whereby one can assume the morphological prefixes/suffixes are relatively frequent. Compound splitting is a non-trivial problem in languages which have a lot of them (notably Swedish and German). An evaluation in terms of improvement in Machine Translation or Information Retrieval is recommended.

Chan, E. (2006). Learning probabilistic paradigms for morphology in a latent class model. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*, pages 69–78. Association for Computational Linguistics, New York City, USA.

Chrupala, G. (2008). *Towards a Machine-Learning Architecture for Lexical Functional Grammar Parsing*. PhD thesis, Dublin City University. Chapter 6.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Hammarström, H. and Borin, L. (2011). Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Kann, K. and Schütze, H. (2016). MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In

Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 62–70. Association for Computational Linguistics, Berlin, Germany.

Khaliq, B. (2015). *Unsupervised Learning of Arabic Non-Concatenative Morphology*. PhD thesis, University of Sussex.

Ma, J., Henrich, V., and Hinrichs, E. (2016). Letter sequence labeling for compound splitting. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 76–81, Berlin, Germany. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119. Neural Information Processing Systems, Lake Tahoe, Nevada.

Moon, T., Erk, K., and Baldridge, J. (2009). Unsupervised morphological segmentation and clustering with document boundaries. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language*

Processing, pages 668–677. Association for Computational Linguistics, Singapore.

Östling, R. and Tiedemann, J. (2017). Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649. Association for Computational Linguistics.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Snyder, B. and Barzilay, R. (2008). Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio. Association for Computational Linguistics.