A Survey and Classification of Methods for (Mostly) Unsupervised Learning of Morphology

Harald Hammarström
Dept. of Computing Science
Chalmers University of Technology
412 96, Gothenburg Sweden
harald2@cs.chalmers.se

Abstract

This paper surveys work on unsupervised learning of morphology. A fairly broad demarcation of the area is given, and a hierarchy of subgoals is established in order to properly characterize each line of work. All the minor and major lines of work are mentioned with a reference and a brief characterization. Different approaches that have been prevalent in the field as a whole are highlighted and critically discussed. The general picture resulting from the survey is that much work has been repeated over and over, with little exchange and evolution of techniques. All in all, the contribution of this paper is a very brief but comprehensive umbrellic synopsis to the research area.

1 Introduction

The problem of (mostly) unsupervised learning of morphology (ULM) may be broadly delineated as follows:

Input: Raw (unannotated) natural language text data

Output: A description of the morphological structure (there are various levels to be distinguished; see below) of the language of the input text

With: As little supervision, i.e. parameters, annotated bootstrapping data, model selection during development etc., as possible

Some approaches have explicit or implicit biases towards certain kinds of languages; they are nevertheless considered to be ULM for this survey.

Morphology may be narrowly taken as to include only derivational and grammatical affixation, where the number of affixations a root may take is finite and the order of affixation may not be permuted. This survey also subsumes attempts that take a broader view including clitics and compounding (and there seems to be no reasons in principle to exclude incorporation and lexical affixes). A lot of, but not all, approaches focus on concatenative morphology/compounding only.

All works in this survey operate on orthographic words – excluding word-segmentation for languages that do not mark word-boundaries orthographically.

One of the matters that varies the most between different authors is the desired outcome. It is useful to set up the implicational hierarchy shown in Table 1 (which need of course not correspond to steps taken in an actual algorithm). The division is implicational in the sense that if one can do the morphological analysis of a lower level in the table, one can also easily produce the analysis of any of the above levels. For example, if one can perform analysis into stem and affixes, one can decide if two word are of the same stem. The converse need not hold, it is perfectly possible to answer the question of whether two words are of the same stem with high accuracy, without having to commit what the actual stem is.

A lot of recent articles do not deal properly with previous and related work, some reinvent heuristics that have been sighted earlier, and there is little modularization taking place. Thus the time is ripe, even

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 292–296

Affix list	A list of the affixes.
↑	
Same-stem decision	Given two words, decide if they are affixations of the same stem.
↑	
Analysis	Given a word, analyze it into stem and affix(es).
↑	
Paradigm list	A list of the paradigms.
↑	
Lexicon+Paradigm	A list of the paradigms and a list of all stems with information of which paradigm each stem belongs to.

Table 1: Levels of power of morphological analysis. We do not make a distinction between probabilistic and non-probabilistic versions.

overdue, for a survey and classification of ideas in this area.

Our full bibliography of ULM-work comprises at least 100 articles/books (more if the level of unsupervised-ness is relaxed out of control) spanning from 1955 to 2006. Clearly, each article cannot be cited or discussed in detail, but we will cover each distinct line of work.

2 Roadmap and Synopsis of Earlier Studies

For reasons of space, very short characterizations of selected representatives of each line of work is given in Table 2. In addition, there is relevant work (Manning, 1998; Borin, 1991; Neuvel and Fulop, 2002) on formalizing morphological regularities but which do not suggest an algorithm that performs on raw text data input.

It was impossible to characterize methods and ideas in brief for each line of work because of the amount of detail necessary to give a relevant comparative picture. However, all work uses some kind of frequency count of n-character grams, and almost all trace their inspiration back to (Harris, 1955). In addition, some recent approches use a Minimum Description Length (MDL)-inspired formula as an optimization criterion of a given model. All the approaches to non-concatenative morphology involve an alignment-step. A few lines of work have tried to exploit other kinds of clues than character sequences, such as similarities in semantics or syntax between words (also acquired in a semi-supervised manner). A fair comparison of previous work in terms of accuracy figures is entirely impossible, not only because of the great variation in goals but also because most descriptions do not specify their algorithm(s) in enough detail. This aspect is better handled in controlled competitions, such as the Unsupervised Morpheme Analysis – Morpho Challenge 2007[1] which a task of segmentation of Finnish, English, German and Turkish.

3 Discussion

Although the heuristic of Harris has had some success it was shown (in various interpretations) as early as (Hafer and Weiss, 1974) that it is not really sound – even for English. In the 2000s, probably independently, a slightly better extension of the same idea emerged, namely, to compile a set of words into a *trie* and predict boundaries at nodes with high actitivity, but this is not sound either as non-morphemic short common character sequences also show significant branching.

So far, all the approaches with mixed MDL-optimization are unsatisfactory on two main accounts; on the theoretical side, they still owe an explanation of why compression or MDL-inspired weighting schemes should give birth to segmentations coinciding with morphemes as linguists conceive of morphemes. On the experimental side, thresholds, supervised/developed parametres and selective input still cloud the success of reported results. What is clear, however, apart from whether it is theoretically motivated, is that MDL approaches are *useful*.

4 Conclusion

What emerges from the last 10 years of intensive research is that, essentially, different people have been doing the same thing with little exchange between each other.

[1] Website `http://www.cis.hut.fi/morphochallenge2007/` accessed 10 January 2007.

	Model	Superv.	Experimentation	Learns what?
(Harris, 1955)+	C	T	English	Analysis
(Andreev, 1965)	C	T	E-type (I)	Unclear
(Lehmann, 1973)	C	T	German	Analysis
(Hafer and Weiss, 1974)	C	T	English	Analysis
(Wothke and Schmidt, 1992)	C	T	German	Analysis
(Klenk, 1994)+	NC	T	Arabic + E-type	Analysis
(Langer, 1991)+	C	T	German	Analysis
(Flenner, 1995)+	C	T	Spanish	Analysis
(Brent et al., 1995)	C	T	English	Analysis
(Džeroski and Erjavec, 2000)	C	T	Slovene	Analysis
(Kazakov and Manandhar, 2001)+	C	T	French/English	Transducer
(Gaussier, 1999)	C	T + AP	English (I)	Paradigms
(Goldsmith, 2006)+	C	T	E-type (I)	Paradigms+Lexicon
(Clark, 2001)+	NC	# states	German/Arabic/English	Transducer
(Déjean, 1998)+	C	T	E-type	Analysis
(Schone, 2001)+	C	T	E-type	Related pairs of words
(Baroni, 2003)+	C	T	E-type	Analysis
(Jacquemin, 1997)	C	T	E-type	Related pairs of words
(Sharma et al., 2002)+	C	T	Assamese	Paradigms+lexicon
(Baroni et al., 2002)	NC	T	English/German (I)	Ranked list of related word pairs
(Creutz, 2006)+	C	T	Finnish/Turkish/English	Analysis
(Kontorovich et al., 2003)	C	T	English	Analysis
(Snover and Brent, 2003)+	C	T	English/Polish	Related pairs of words
(Johnson and Martin, 2003)	C	T	Inuktitut	Unclear
(Wicentowski, 2004)+	NC	AP	30-ish E-type	Transducers
(Ćavar et al., 2004)+	C	T	Unclear	Paradigms
(Argamon et al., 2004)	C	T	English	Analysis
(Goldsmith et al., 2005)+	NC	T	Unclear	Unclear
(Oliver, 2004, Ch. 4-5)	C	T	Catalan	Paradigms
(Kurimo et al., 2005)	C	T	Finnish/Turkish/English	Analysis
(Hammarström, 2006)+	C	-	Maori to Warlpiri	Same-stem

Table 2: Very brief roadmap of earlier studies. Abbrevations in the Table: C = Concatenative, NC = Also non-concatenative, T = Thresholds and Parameters to be set by a human, AP = Aligned pairs of words, E-type = European Indo-European type languages, I = Impressionistic evaluation. + = entry also covers earlier work by the same author(s).

References

Nikolai Dmitrievich Andreev, editor. 1965. *Statistiko-kombinatornoe modelirovanie iazykov*. Akademia Nauk SSSR, Moskva.

Shlomo Argamon, Navot Akiva, Amihood Amit, and Oren Kapah. 2004. Efficient unsupervised recursive word segmentation using minimum description length. In *COLING-04, 22-29 August 2004, Geneva, Switzerland*.

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL/SIGPHON-2002*, pages 48–57.

Marco Baroni. 2003. Distribution-driven morpheme discovery: A computational/experimental study. *Yearbook of Morphology*, pages 213–248.

Lars Borin. 1991. *The Automatic Induction of Morphological Regularities*. Ph.D. thesis, University of Uppsala.

Michael R. Brent, S. Murthy, and A. Lundberg. 1995. Discovering morphemic suffixes: A case study in minimum description length induction. In *Fifth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale, Florida*.

Damir Ćavar, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, and Giancarlo Schrementi. 2004. On induction of morphology grammars and its role in bootstrapping. In Gerhard Jäger, Paola Monachesi, Gerald Penn, and Shuly Wintner, editors, *Proceedings of Formal Grammar 2004*, pages 47–62.

Alexander Clark. 2001. Partially supervised learning of morphology with stochastic transducers. In *Proc. of Natural Language Processing Pacific Rim Symposium, NLPRS 2001*, pages 341–348, Tokyo, Japan, November.

Mathias Creutz. 2006. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition*. Ph.D. thesis, Helsinki University of Technology, Espoo, Finland.

Hervé Déjean. 1998. *Concepts et algorithmes pour la découverte des structures formelles des langues*. Ph.D. thesis, Université de Caen Basse Normandie.

Sašo Džeroski and Tomaž Erjavec. 2000. Learning to lemmatise slovene words. In James Cussens and Saso Dzeroski, editors, *Learning Language in Logic*, volume 1925 of *Lecture Notes in Computer Science*, pages 69–88. Springer-Verlag, Berlin.

Gudrun Flenner. 1995. Quantitative morphsegmentierung im spanischen auf phonologischer basis. *Sprache und Datenverarbeitung*, 19(2):63–78.

Éric Gaussier. 1999. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999)*. Association for Computational Linguistics, Philadephia.

John Goldsmith, Yu Hu, Irina Matveeva, and Colin Sprague. 2005. A heuristic for morpheme discovery based on string edit distance. Techincal Report of Computer Science Department, University of Chicago.

John A. Goldsmith. 2006. An algorithm for the unsupervised learning of morphology. *Computational Linguistics*, 12(4):353–371.

Margaret A. Hafer and Stephen F. Weiss. 1974. Word segmentation by letter successor varieties. *Information and Storge Retrieval*, 10:371–385.

Harald Hammarström. 2006. Poor man's stemming: Unsupervised recognition of same-stem words. In Hwee Tou Ng, Mun-Kew Leong, Min-Yen Kan, and Donghong Ji, editors, *Information Retrieval Technology: Proceedings of the Third Asia Information retrieval Symposium, AIRS 2006, Singapore, October 2006*, volume 4182 of *Lecture Notes in Computer Science*, pages 323–337. Springer-Verlag, Berlin.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Christian Jacquemin. 1997. Guessing morphology from terms and corpora. In *Proceedings, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97), Philadelphia, PA*.

Howard Johnson and Joel Martin. 2003. Unsupervised learning of morphology for English and Inuktitut. In *HLT-NAACL 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, May 27 - June 1, Edmonton, Canada*, volume Companion Volume - Short papers.

Dimitar Kazakov and Suresh Manandhar. 2001. Unsupervised learning of word segmentation rules with genetic algorithms and inductive logic programming. *Machine Learning*, 43:121–162.

Ursula Klenk. 1994. Automatische morphologische analyse arabischer wortformen. In Ursula Klenk, editor, *Computatio Linguae II: Aufsätze zur algorithmischen und Quantitativen Analyse der Sprache*, volume 83 of *Zeitschrift für Dialektologie und Linguistik: Beihefte*, pages 84–101. Franz Steiner, Stuttgart.

L. Kontorovich, D. Don, and Y. Singer. 2003. A markov model for the acquisition of morphological structure. Technical report, CMU-CS-03-147, School of Computer Science, Carnegie Mellon University, June.

Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy, and Murat Saraclar. 2005. An introduction and evaluation report. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Unsupervised segmentation of words into morphemes – Challenge 2005*.

Hagen Langer. 1991. *Ein automatisches Morphsegmentierungsverfahren für deutsche Wortformen*. Ph.D. thesis, Georg-August-Universität zu Göttingen.

Hubert Lehmann. 1973. *Linguistische Modellbildung und Methodologie*. Max Niemeyer Verlag, Tübingen. Pp. 71-76 and 88-93.

Christopher D. Manning. 1998. The segmentation problem in morphology learning. In Jill Burstein and Claudia Leacock, editors, *Proceedings of the Joint Conference on New Methods in Language Processing and Computational Language Learning*, pages 299–305. Association for Computational Linguistics, Somerset, New Jersey.

Sylvain Neuvel and Sean A. Fulop. 2002. Unsupervised learning of morphology without morphemes. In *Workshop on Morphological and Phonological Learning at Association for Computational Linguistics 40th Anniversary Meeting (ACL-02), July 6-12*, pages 9–15. ACL Publications.

A. Oliver. 2004. *Adquisició d'informació lèxica i morfosintàctica a partir de corpus sense anotar: aplicació al rus i al croat*. Ph.D. thesis, Universitat de Barcelona.

Patrick Schone. 2001. *Toward Knowledge-Free Induction of Machine-Readable Dictionaries*. Ph.D. thesis, University of Colorado.

Utpal Sharma, Jugal Kalita, and Rajib Das. 2002. Unsupervised learning of morphology for building lexicon for a highly inflectional language. In *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON), Philadelphia, July 2002*, pages 1–10. Association for Computational Linguistics.

Matthew G. Snover and Michael R. Brent. 2003. A probabilistic model for learning concatenative morphology. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1513–1520. MIT Press, Cambridge, MA.

Richard Wicentowski. 2004. Multilingual noise-robust supervised morphological analysis using the wordframe model. In *Proceedings of the ACL Special Interest Group on Computational Phonology (SIGPHON)*, pages 70–77.

Klaus Wothke and Rudolf Schmidt. 1992. A morphological segmentation procedure for german. *Sprache und Datenverarbeitung*, 16(1):15–28.