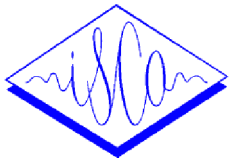

ITRW on Experimental
Linguisriucs (ExLing 2008)
Athens, Greece
August 25-27, 2008

MORPHEMIA: a semi-supervised algorithm for the segmentation of Modern Greek words into morphemes

Constandinos Kalimeris and Stelios Bakamidis
Voice and Sound Technology Department, Institute for Language and Speech Processing (ILSP), Greece

Abstract

The present paper reports on MORPHEMIA, a semi-supervised machine-learning algorithm designed to segment Modern Greek (MG) words into morphemes. The algorithm segments its input iteratively. During its first iteration, the algorithm uses its a priori linguistic knowledge. At the end of each successful iteration, the algorithm extracts new morphological knowledge which is utilised during its next iteration. Thus, with each successful iteration, the algorithm segments an increasing amount of its input data. The algorithm uses a metric to decide whether a given extracted piece of morphological knowledge will improve its performance and only accepts it if it will. Thus, its output gradually improves in quality. MORPHEMIA terminates its operation when new knowledge can no longer be extracted from its input data.

Key words: machine-learning, morphologically rich languages, statistical language modeling, morphological lexicons

Introduction

Recently, the idea of building statistical language models based on linguistic units smaller than the word has been gaining ground (Kirchhoff and Sarikaya 2007). The advantages of utilising a unit such as the morpheme in language technology applications handling linguistic data of Morphologically Rich Languages (MLRs) may seem obvious. However, morphemes are notoriously difficult to define and demarcate within words. MORPHEMIA attempts to bridge the gap between two radically different approaches to the task of the morphological segmentation of words: the time-consuming and expensive manual segmentation by linguists; and the completely unsupervised but theoretically blind automatic segmentation by language engineers, who try to bypass the problem by utilizing unsupervised machine-learning algorithms. The latter perform the task without the help of any a priori linguistic knowledge, extracting knowledge directly from the data (Siivola et al. 2007).

Proceedings of the 2nd ISCA Workshop on Experimental Linguistics, ExLing 2008, 25-27 August 2008, Athens, Greece.

Description

A priori linguistic knowledge

MORPHEMIA uses two types of linguistic knowledge: lists of re-write rules for morphemes and a manually segmented sample of its lexical input.

MORPHEMIA uses five lists (L1 – L5) which contain re-write rules pertinent to five different types of morphemes and morpheme clusters that we assume to be relevant to the morphology of Modern Greek (MG). L1 contains "exceptions", mainly rules for whole words whose morphological structure conforms to rare patterns, usually because of their ancient or foreign origin (these are often monosyllable and/or function words.) L2 contains rules for morphemes or morpheme clusters which can occur at the end of MG words and which may or may not coincide with the traditional "endings" listed in grammars of MG (e.g. ω→-ώ, ός→-ός, είς→-είς, ικός→-ικ-ός, άδες→-άδ-ες, etc, the hyphens implying morpheme boundaries). L3 contains rules pertinent to traditional "roots" or clusters of them commonly occurring in MG compounds. (The latter can include derivational morphemes,e.g. ολογ→-ο-λογ-, οικονομ→-οικ-ο-νομ-, ικοποίησ→-ικ-ο-ποί-ησ-, etc.) L4 contains rules for (clusters of) derivational morphemes occurring to the right of "roots" (e.g. ατισμέν→-ατ-ισ-μέν-, εύθηκ→-εύ-θ-ηκ-, ήσιμ→-ήσ-ιμ-, ότητ→-ότ-ητ-, etc). Finally, L5 contains rules for (clusters of) derivational morphemes occurring to the left of "roots" and/or at the beginning of words (αντι→-αντι-, ξανα→ξανα-, επανα→επ-ανα-, αναδια→-ανα-δια-, etc).

Prior to its operation, MORPHEMIA needs to be supplied with a random and representative sample of the wordlist it will process. The words of the sample must be manually segmented into morphemes. The segmented sample will then function as MORPHEMIA's "Golden Standard". In essence, the manually segmented sample is assumed to be an implicit (yet adequate) statement of MG morphology. While creating the Golden Standard, the users of MORPHEMIA may introduce as much or as little linguistic knowledge as possible or desirable.

Operation

The operation of MORPHEMIA is outlined in Figure 1. During its first iteration, MORPHEMIA uses the knowledge contained in L1 – L5 to segment the *sample*. This automatically segmented sample is compared against the Golden Standard. The metric *b* expresses the precision of the first segmentation: *b* = (correctly segmented words / N) X 100, where N is the number of words comprising the sample.

Next, MORPHEMIA searches within the words of its entire *input* for the "target strings" contained in lists L1 – L5 (i.e. for the strings to the left of the rules' arrows) and replaces them with the dummy character "*".

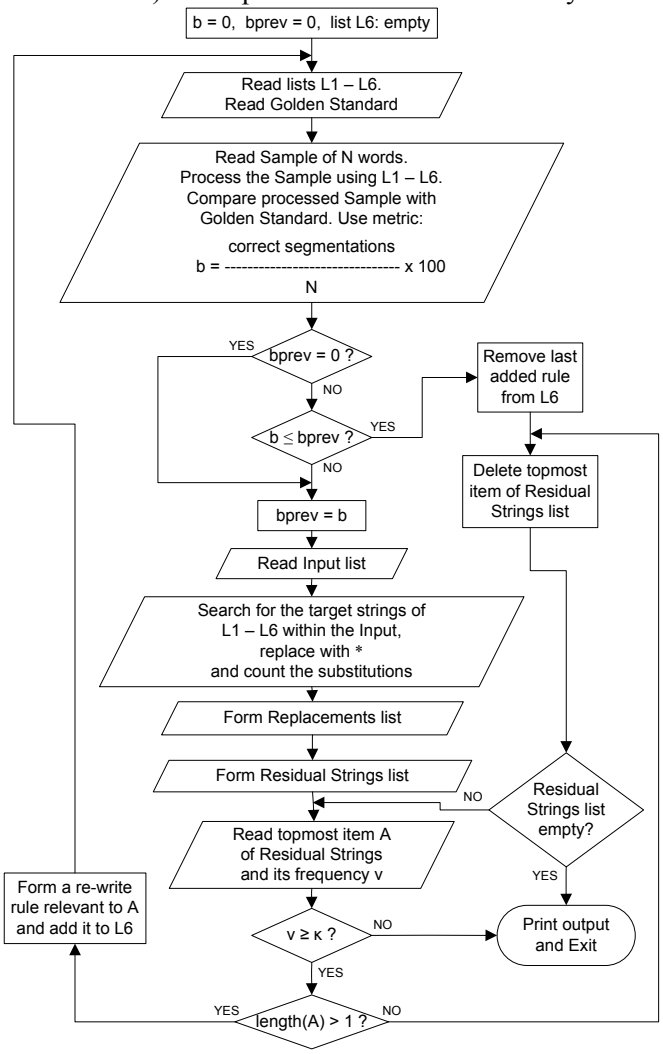

Figure 1. Flowchart of the MORPHEMIA algorithm.

Following that, MORPHEMIA counts the frequency of the substrings which were not replaced by "*" and creates the Residual Strings list.

The most frequent of these sub-word units, "A", is then tested for its validity as a legitimate MG morpheme. A relevant rule (A→-A-) is formed and added to the as yet empty list L6. Then, MORPHEMIA re-segments the sample using the contents of L1 – L5 but also of L6. If the new *b* metric is

smaller than or equal to the previous one, the rule is removed from L6, the relevant string is removed from the Residual Strings list and the algorithm repeats the process with the next most frequent residual string. However, if the new b is greater than the previous one, the string is accepted as a legitimate morpheme and the relevant rule remains in L6, thus becoming part of the a priori knowledge to be used in the next iteration. The user may define a floor value κ as the lowest frequency which determines if a substring will be considered at all (in our experiments, $\kappa = 1$). Note that, although one-character long morphemes can be justified in MG morphology, MORPHEMIA disallows them to avoid over-segmentation of the input.

MORPHEMIA continues its operation until the Residual Strings list is empty. The contents of L6 represent the new morphological knowledge extracted by the algorithm.

Brief discussion of experimental results

A number of experiments have been conducted, using an input list of 226,857 different words, a Golden Standard (GS) of N=1000, and a Final Evaluation Sample (FES) of N=1000, different from GS. For lack of space here, we report on MORPHEMIA's worst and best performances. ***Worst:*** a priori knowlegde: 69 rules (L1: 0, L2: 39, L3: 6, L4: 7, L5: 17), $b_{\text{GS-initial}}$: 11.70%, $b_{\text{GS-final}}$: 41.70%, input characters affected: 56.53%, L6: 185 rules, $b_{\text{FES-final}}$: 25.20%. ***Best:*** a priori knowlegde: 2527 rules (L1: 22, L2: 508, L3: 1552, L4: 258, L5: 187), $b_{\text{GS-initial}}$: 53.70%, $b_{\text{GS-final}}$: 74.50%, input characters affected: 75.22%, L6: 127 rules, $b_{\text{FES-final}}$: 60.10%.

As expected, MORPHEMIA's performance depends on the amount of a priori knowledge, especially on the size of L3 (rules for "roots"). Interestingly, the candidate morphemes *rejected* by the b metric evaluation process are a valuable source of new candidate roots. These can be manually selected and added to the knowledge included in L3. Thus, MORPHEMIA can also prove a valuable tool for building morphological lexicons: the ratio "input words / strings considered" is roughly 11/1.

References

Kirchhoff, K. and Sarikaya, R. 2007. Processing morphologically rich languages. Research Tutorial at the 8th Annual Conference of the International Speech Communication Association, Interspeech 2007. Antwerp, Belgium.

Siivola, V., Creutz, M. and Kurimo, M. 2007. Morfessor and VariKN machine learning tools for speech and language Technology. Proceedings of the 8th Annual Conference of the International Speech Communication Association, Interspeech 2007. Antwerp, Belgium.