Unsupervised Learning of Naïve Morphology with Genetic Algorithms

Dimitar Kazakov

Department of Control Engineering
Czech Technical University, Prague
http://labe.felk.cvut.cz/ai_group/kazakov/
kazakov@labe.felk.cvut.cz

Abstract

The morphological lexicon is an important part of NLP systems which is typically hand-written with the help of linguist experts. Even a partial automation of this process could decrease the cost of the lexicon, being of theoretical importance for languages and dialects which have not been well analysed yet. In this work we describe an attempt to use the minimal description length (MDL) as the one bias for deriving lexicons of morphemes from a raw list of words. MDL is used as a fitness function of a simple genetic algorithm. Results are reported for a rich-morphology language corpus (French) and future work is discussed.

1 Introduction

“A struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand.”

Charles Darwin: The Descent of Man, and Selection in Relation to Sex.

This paper presents unsupervised learning of word morphology from a raw list of words. The approach uses a genetic algorithm with a minimal description length bias. Segmentation of the given words into morphemes is obtained as a result, along with pairs of formative lexicons.

Let us see what could motivate the use of a learning approach in a field where a fair number of morphological dictionary acquisition systems have already been developed.
The morphological dictionary acquisition is usually divided between a linguist, specifying the framework and a lexicographer adding entries. The morphemes induced by our algorithm could be used after a hand-made verification in the language model. Also boundaries between word morphemes are proposed to the lexicographer. However, the theoretical possibility to summarise the linguistic knowledge about word formation and replace it with the simple principle of minimal description length, even as a first approximation, is the principal question which has been brought up in this work.

2 Morphology—The Classic Approach

The standard linguistic approach is to see the word formation as a concatenation of morphemes – *prefixes, stems, suffixes and endings*. There are different theories, some of which considering a higher number of invariable morphemes, others making use of transformations on the boundaries between morphemes in order to reduce the size of morpheme lexicons. For instance, the Latin verb paradigm can be described in one of the three following ways [Ben]:

1. Conjugational solution: invariant stems and variant suffixes, e.g., *am-a: mus, mon-e: mus, teg-i: mus, aud-i: mus, am-ant, mon-ent, teg-unt, aud-iunt*. It is one type of Item and Arrangement (IA) solution [Hoc54], one that consists of items (the stems and the suffixes) and their arrangements (an indication of which suffix variants go with which stems.) This is the common way to show the verb conjugation in the dictionaries.

2. Variant stem solution: invariant suffixes and variant stems – *ama:-mus, mone:- mus, tegi:-mus, audi:-mus, ama-nt, mone-nt, tegu-nt, audiu-nt*. It is also an IA solution. The segmentation is done further to the right, so that the varying vowel is included in the stems – the variability is transferred to the stems. This solution allows some very general statements, for instance, to say that the verb form for the first person of plural in French terminates by *-ons: Nous plantons, nous suivrons, nous rendrons* [GG89].

3. Phonological solution: it is possible to consider both stems and invariant suffixes as invariable and use rules to adjust the morphemes on their boundaries in order to fit together. It is similar to what was proposed by [Hoc54] as Item and Process solution—the items being the stems and suffixes, and the processes the necessary rules. For instance, *ama-, mone-, teg-, aud-* could be the list of stems and *-mus, -nt* the list of suffixes. Incorrect combinations, such as *mone:nt* can be handled by rules of the sort *Shorten long vowels before final -nt*.

It is obvious that the “correct” morphological analysis of a word varies according to the chosen approach and that it is difficult to favour one over another.
When discussing morphology, an important distinction is made between word inflection and word derivation. The former corresponds to word forms of the same paradigm, keeping the same part of speech and the same meaning. Only morphological categories such as number and gender are changed. Word derivation, even if referring to the general semantic value of the starting word, modifies it and the resulting word can correspond to a different part of speech. A good illustration of word derivation is, for instance, the couples adjective–abstract noun: good–goodness etc.

The complexity of word morphology varies in large boundaries between different languages. The languages which use mainly word inflection to denote the word role in the sentence are called synthetic languages\(^1\). These languages have the most complex word morphology. The paradigm of Czech adjectives and nouns, for instance, includes word forms for seven cases, four genders\(^2\) and three numbers\(^3\). Analytic language syntax, on the other hand, uses combinations of words rather than different word forms. In general, analytic languages have a smaller number of word forms and no declension. However, the verb paradigm could be very complex and can consist of tens of word forms as in French or in the other Roman languages.

3 Naïve Theory of Morphology and MDL Bias

Even if the classic theory of morphology involves several types of morphemes, some of which are able to appear in the word more than once, such an approach has some disadvantages in the natural language processing field. It either produces a lot of spurious analyses or requires the use of some additional constraints, which slows down the algorithm and increases the cost of the system development.

For instance, in some language families new verbs can be derived from the basic stem by prefixing, the English examples being undo and co-operate. Either all the items do, undo, operate, co-operate are enumerated in the stem lexicon, or separate prefix (un-, co-) and stem (do, operate) lexicons are introduced. The simple statement \(<\text{Verb}:><\text{Prefix}><\text{Stem}>\) gives too much freedom to the system, e.g., the word unknown would be seen as the past participle of the verb unknown. An alternative could be the application of complex rules involving word semantics.

We will accept here that the number of word morphemes is limited to two per lexical unit. They will be called further ‘stems’, resp. ‘suffixes’. It will be shown further how the limitations of the assumption that we make can be overcome by the recurrent use of the same approach.

---

\(^1\)Latin, German, almost all Slavonic languages

\(^2\)animate masculine, inanimate masculine, feminine, neuter

\(^3\)singular, dual and plural

107
For a given set of words the positions of morpheme boundaries correspond to a vector of integers, each of them between one and the word length. The word set along with an arbitrary vector of that kind introduces a naïve morphological theory. Every such theory defines two lexicons where stems, resp. suffixes are enumerated without repetition (see fig.1).

The quality of the theory will be estimated by the number of characters $N$ that both lexicons contain—the smaller that number, the better the theory. The upper bound $N_{\text{max}}$ of that measure is given by the number of characters in the word list. This case is reached when no stem nor suffix has been generated twice by the theory.

![Table showing word lengths, morpheme boundaries, and lexicons]

**Figure 1**: A naïve theory of word morphology

The quality criterion is based on the hypothesis that substrings composed out of real morphemes occur in the words with a frequency higher than any other left or right substrings. In that way, a theory with a low $N$ would produce lexicons where 'stems' and 'suffixes' correspond very often to single morphemes or their combinations. Since the word list can be stored as a list of couples of indices $<\text{stem}, \text{suffix}>$ along with the lexicons of stems and suffixes, the search for a theory minimising $N$ can be seen as a *minimal description length* (MDL) task.

### 4 Mapping The MDL Search To GA

A large search space and a lack of applied algorithms are good premises for the use of genetic algorithms (GA). The used approach is rather straight-forward, making use of a simple genetic algorithm [Gol89], where the vector of integers corresponding to the theory is directly used as an individual (chromosome) and its fitness function is $N_{\text{max}} - N$. The fitness function defined in that way does not take negative values and reformulates the MDL task to the conventional for a GA search of individuals with maximal fitness. In the context of individuals with non-binary features, mutation is
defined either as a shift of the morpheme boundary with one position to the left or right, or, as a random choice of a new boundary position from the allowed interval. The best ever individual is kept through the generations and is finally given as a result of the learning.

5 Results

A list containing many forms of the same word paradigm would be suitable for the learning since the difference between $N_{max}$ and $N$ of the best solution is very great. An easy way to obtain such a word list still remaining within the framework of unsupervised learning is to sort the words of a large corpus and remove all multiple occurrences of the same word forms. With the increasing corpus size, the word list converges on an exhaustive lexicon of word forms. Parts of an already available French lexicon of that kind were actually used for our experiments.

The computational complexity of the algorithm limits the size of the input lists to hundreds of words. However, a large amount of input data is not of such importance for the results, as far as many instances of the same paradigm are included. A reference experiment was carried out, when the input only consisted of 39 forms of the same verb ($\text{annonner}$). The derived paradigm had five stems. Apart from $\text{annonna}$-, all other stems were typical for some verb mood or tense: $\text{annon}$- (present tense), $\text{annonass}$- (subjunctive), $\text{annonner}$- (infinitive, imperfect, future tense), $\text{annonne}$- (past participle). The lexicon of suffixes contained 21 items. They are enumerated below, grouped according to some of the verb tenses or moods they correspond to: $-ai$, $-as$, $-a$, $-ons$, $-ez$, $-ont$ (future tense), $-ais$, $-aût$, $-ions$, $-iez$, $-aient$ (imperfect), $-e$, $-es$, $-ent$ (present tense), $-èmes$, $-êtes$, $-érent$ (simple past tense), $-ât$ (subjunctive), $-s$ (plural of the past participle). The last two suffixes, $-nt$ and $-sse$ are the shortened forms of the present participle suffix $-ant$, resp. the subjunctive imperfect tense suffix $-asse$.

A higher number of morphemes, e.g., three in the word, would influence unfavourably the computational complexity of the task. Moreover, any fixed value does not reflect the fact that the number of morphemes cannot be unified for all words. A more feasible way to divide the found morpheme patterns into smaller parts is the recurrent use of the same algorithm with each of the derived lexicons as an input. Continuing with our example, the learning from the list of five stems resulted in the correct invariable stem $\text{annon}$- and four “intermediate” suffixes $-er$, $-ass$, $-a$, $-é$.

There is one typical example of the tests we carried out. The word list contained 120 items, the number of individuals in the population was set to 800. The word list contained 92 verbs (forms), 16 nouns, 9 adjectives, one preposition, one pronoun and one adverb. The average time for the computing of one generation was 15 seconds on
Pentium 90 Mhz platform. The starting list contained words with average length 7.7 \( (N_{\text{max}} = 922) \). The naïve theory learned after 2000 generations and almost eight and half hours had \( N = 304 \), i.e. \( N/N_{\text{max}} = 0.33 \). For comparison, the hand-made morphological analysis which corresponds to the conjugational (constant stem) paradigm has the ratio \( N_{\text{standard}}/N_{\text{max}} = 0.38 \). A sample of the resulting word segmentation is shown in tab.1.

\[
\begin{align*}
\text{ânon+onna} & \quad \text{ânonnass+ses} & \quad \ldots \\
\text{ânonn+i} & \quad \text{ânonnass+iez} & \quad \text{ânonn+âmes} \\
\text{ânonn+aient} & \quad \text{ânonnass+ions} & \quad \text{ânonn+ât} \\
\text{ânonn+is} & \quad \text{ânonn+e} & \quad \text{ânonn+âtes} \\
\text{ânonn+ait} & \quad \text{ânonne+nt} & \quad \text{ânonn+èrent} \\
\text{ânonn+ant} & \quad \text{ânonn+er} & \quad \text{ânonn+é} \\
\text{ânonn+s} & \quad \text{ânonne+ra} & \quad \text{ânonn+ée} \\
\text{ânonn+sse} & \quad \text{ânonnera+i} & \quad \text{ânonn+ées} \\
\text{ânonn+ssent} & \quad \text{ânonner+aien} & \quad \text{ânonn+és}
\end{align*}
\]

Table 1: A sample of learned naïve theory

The series of tests pointed out some interesting features in the algorithm behaviour. The results are closer to the variable stem solution, shortening the suffixes. Suffixes are merged inside the paradigm (ânonn/ânonnass+ions rather than ânonn+ions/assions) as well as between similar paradigms (ânonn/êbahì+sses).

The comparative tests between the two mutation techniques described in the previous section showed a surprising result. Even if the “random position” mutation guaranteed the population higher average fitness, the best final results (best-ever individuals) were reached with the “shift” mutation. The latter technique was used in the rest of the tests.

Nondeterminism is characteristic of GA, so that several runs with the same data result in different naïve theories. A certain amount of noise can be removed by simple merging of these theories and counting for all of them the average frequency of each morpheme.

An overview of the average results of 13 runs is displayed in tab.2. In the first part of the table, all the word forms from the input list are represented by their basis forms in which they appear in the dictionaries. Only for the purpose of evaluation, each word basis form is also matched to the corresponding part of speech (POS)—this information has not been used during the learning phase. The merged naïve theory considered 85 different stems and 215 suffixes. However, only 30 stems and 29 suffixes occurred in average once or more. They are shown in the second, resp. the third part of tab.2 in decreasing order of average frequency.

With regard to the different theoretical approaches described in section 2, it is difficult to give an exact analytic evaluation of the acquired results. The most important result is the patterns, both stems and suffixes that are learned. Their occurrence as a left,
Table 2: Lists of the word basis forms, derived stems and suffixes

<table>
<thead>
<tr>
<th>Words</th>
<th>POS</th>
<th>Stems</th>
<th>Suffixes</th>
</tr>
</thead>
<tbody>
<tr>
<td>acre</td>
<td>Adj</td>
<td>13.3</td>
<td>2.0</td>
</tr>
<tr>
<td>acrétè</td>
<td>N</td>
<td>7.8</td>
<td>ébalhí</td>
</tr>
<tr>
<td>age</td>
<td>N</td>
<td>7.3</td>
<td>ébarb</td>
</tr>
<tr>
<td>agé</td>
<td>Adj</td>
<td>7.3</td>
<td>ébalhir</td>
</tr>
<tr>
<td>âme</td>
<td>N</td>
<td>7.2</td>
<td>ébalhiss</td>
</tr>
<tr>
<td>âne</td>
<td>N</td>
<td>6.4</td>
<td>ânonner</td>
</tr>
<tr>
<td>ânerie</td>
<td>N</td>
<td>5.2</td>
<td>ânonna</td>
</tr>
<tr>
<td>ânesse</td>
<td>N</td>
<td>4.7</td>
<td>ébarber</td>
</tr>
<tr>
<td>ânonner</td>
<td>V</td>
<td>4.3</td>
<td>ânonne</td>
</tr>
<tr>
<td>âpre</td>
<td>Adj</td>
<td>4.0</td>
<td>ébarba</td>
</tr>
<tr>
<td>âtre</td>
<td>N</td>
<td>3.2</td>
<td>ébahí</td>
</tr>
<tr>
<td>ère</td>
<td>N</td>
<td>2.6</td>
<td>ânonn</td>
</tr>
<tr>
<td>ès</td>
<td>Prep</td>
<td>2.4</td>
<td>âg</td>
</tr>
<tr>
<td>ça</td>
<td>Pron</td>
<td>2.3</td>
<td>âne</td>
</tr>
<tr>
<td>çà</td>
<td>Adv</td>
<td>2.3</td>
<td>âgé</td>
</tr>
<tr>
<td>ébahir</td>
<td>V</td>
<td>2.2</td>
<td>ân</td>
</tr>
<tr>
<td>ébarber</td>
<td>V</td>
<td>2.0</td>
<td>ébarbe</td>
</tr>
</tbody>
</table>

resp. right substring of a word denote a possible boundary between morphemes. The proposed morphemes are not equally reliable. The experimental results confirmed that, in general, the more frequently an item appears in the derived theory, the higher is the probability that the item is really a morpheme. For instance, 11 of the first 16 candidates could be considered as productive stems. Also all of the top 14 suffixes correspond to some of the invariable stem paradigms.

6 Conclusion

French is an analytic language with a complex verb paradigm, but also adjectives and nouns varying in number and gender. The search space of the learning task is given by the number of possible word divisions into two parts—equal to the number of word letters (an empty stem is not allowed) multiplied by the number of words, i.e., \( l_1 l_2 \ldots l_n \) where \( l_i \) is the length of word \( i \). This can also be approximated by \( (l_{av})^n \), \( l_{av} \) being the average word length and \( n \) the number of words in the list. It is evident that the search space is extremely large. Before we start looking for the most effective algorithm, the MDL bias should have been tested. The GA served as an easy test platform. The results attained confirmed that the postulated hypothesis demands more attention. Although the MDL bias allows to discover some strong regularities in the data, the result at the level of individual words might be improved by a further refining. One of the promising ways leads through the choice of a word decomposition into mor-
phemes which have a high frequency. The approach described in this paper could be additionally improved in two almost independent ways. First, the effectiveness of the algorithm could be influenced by the application of some advanced genetic operators. There are also some perspective strategies related to the specific learning task. It can be observed from tab.2 that the algorithm is “hesitating” between the conjunctional solution (shorter stems) and variable stem solution (shorter suffixes). It is possible that giving one of the lexicons a relatively higher weight would lead to shorter morphemes of the corresponding type. Nevertheless, such *ad hoc* parameters would be a considerable theoretical weakness of the approach.

The MDL bias which is currently used does not assign higher weight to either morpheme lexicon. That balance is upset by the input list alphabetical sorting. The words could be also sorted in reverse order, so that the input list would contain several items with the same suffix. Such an experiment has not been run yet. However, comparison between the results of the learning for the alphabetical and reversed word order is planned in our future work.

References


