Morphology based automatic acquisition of large-coverage lexica

Lionel Clément, Benoît Sagot, Bernard Lang

INRIA – Institut National de Recherche en Informatique et en Automatique
Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay, France
{lionel.clement, benoit.sagot, bernard.lang}@inria.fr

Abstract

In this article, we introduce a new technique for constructing wide-coverage morphological lexica from large corpora and morphological knowledge, with an application to French. Basically, it relies on the idea that the existence of a hypothetical lemma can be guessed if several different words found in the corpus are best interpreted as morphological variants of this lemma. We first validated our technique by extracting verbs and adjectives on a general French corpus of 25 million words. Compared with other lexical resources available for French, our results are very satisfying, since we cover many words, often derived words, that are not always present in other lexica. Application of our algorithm to the acquisition of domain-specific adjectives on a botanic corpus gave also very good results, thus demonstrating its usability to extract domain-specific lexica. Moreover, it is generalizable to any language with a substantial morphology. Part of the resulting lexicon (currently verbal forms) is already freely available on `http://www.lefff.net/`.

1. Introduction

Language processing tasks such as wide-coverage parsers for natural language are often limited because of the lack of lexical resources. Corpus processing for lexical acquisition is now widely used. But lexical acquisition is mostly targeted towards advanced information such as terminology (Daille, 2000), collocations (Dunning, 1993), or sub-categorization properties (Briscoe and Carroll, 1997), and assume the availability of basic information such as parts-of-speech and inflectional categories, usually referred to as lexica. However, the acquisition of this basic information gets relatively little discussion in the literature, partly because this problem is both minimal and largely solved for English, and although it is a prerequisite to the acquisition of more advanced information. For languages like French, large-coverage lexical resources are not numerous, and are not always freely available.

Our approach makes it possible to extract from a large raw corpus a list of lemmas with their associated morphological information (part-of-speech, inflectional category, as well as prefix when appropriate). The resulting lexicon can then be compared with other available lexica, usually commercial or with restricted licenses, such as the morphological word lists extractable from ATILF databases (Dendien and Pierrel, 2003), MulText (Ide and Véronis, 1994), or ABU (Association des Bibliophiles Universels, ABU). These lexica are mostly built by human lexicographers, and therefore are both costly to develop and subject to mistakes and omissions. The results of our work is complementary to these lexica, since such problems can be partially avoided by the mechanization of the process, although it has other drawbacks, like over-generation. Since our method is based on corpus statistics, it is related to works on terminology acquisition. However, our goal is not to acquire (possibly multi-words) terms, but lexical entries.

2. General framework

The acquisition of the lexicon of a given corpus can be seen as the extraction of the lemmas of all forms of the corpus among all morphologically acceptable lemmas which have at least one of their form that is attested in the corpus. This extraction is then a matter of separating correct lemmas from incorrect ones among all possible lemmas. To achieve this separation, one need a way to grade possible lemmas according to their plausibility.

At that point, the acquisition process described in this article is based on the following underlying idea. The plausibility of a lemma is correlated with the number of different forms of lemma attested in the corpus, and — to a lesser extent — with the number of occurrences of these forms. This supposes that the language studied has a morphology that is rich enough to associate several forms to each lemma. It is the case for example for romance languages (especially for verbal inflection), but also for rich-morphology languages like for example slavic languages or latin.

In French, for example, the richest morphology is that of verbs. A typical verbal inflection contains more than 40 different forms. Thus, if attested forms constitute a significant part of the inflection of a possible verbal lemma, this lemma is most probably correct, and can be acquired.

3. Algorithm overview

Before reaching the core loop of our algorithm, there are two preliminary steps to go through.

The first preliminary step of our algorithm is to tokenize the corpus. In the following, we shall use the word *form* as a synonym for *token*. We call *tagged form* a form supplemented with morphological information such as gender, number, tense, person, etc. An inflectional category is an operator that computes a set of tagged forms from a single input called *canonical form*. A *lemma* is defined as a canonical form associated with an inflectional category. Our aim is to acquire a full-forms lexicon that links tagged forms to their lemma(s). To achieve the tokenization of our corpus, we used Clement's tokenizer `splitwords`, simply based on regular expressions.

Once tokenized, the corpus is tagged to split up forms according to their part-of-speech. This step is not absolutely necessary to acquire the most frequent lexical units. However, to achieve quasi-exhaustivity, we have to associate to rare forms their part-of-speech. Indeed, the part-

of-speech of a lemma (and thus its inflectional category) is almost impossible to discover by morphological analysis alone, if only a very low number of different inflected forms represent it in the corpus. But this doesn't mean that we extract the lexicon on which the tagger relies — if any. Indeed, the tagger we used, namely `TreeTagger`, has been trained on a 700,000-word tagged corpus and with a lexicon that includes only closed-class words (articles, pronouns, prepositions, conjunctions, etc.), but no open-class word (noun, adjective, verb, adverb). Any other tagger not relying on an open-class lexicon could have been used to perform this tagging task[1].

Our extraction of the lexicon from the corpus can then be seen as the iteration of a three steps loop:

- Construction of possible lemmas,
- Ranking of these lemmas,
- Hand validation of best-ranked not already validated lemmas.

3.1. Construction of possible lemmas

This step can be divided in three parts. Firstly, and if we are not in the first execution of the loop, information given by hand during previous steps (in a way described hereafter, see 3.3.) are used to filter the corpus to reject misspelled words and other typos.

The second part is the construction of possible lemmas itself. This strongly relies on the availability of morphological knowledge. We have developed an inflection module for french verbs, adjectives and nouns (adverbs being invariable), based on standard descriptive grammars for French language (Bescherelle, 1990; Grevisse, 1993). For each not previously filtered word of the corpus, we use this module backwards to compute all possible lemmas having this word as a morphological variant. Our morphological module is precise enough to generate only lemmas that match French morphology. We then filter the result of this process by ruling out the (few) lemmas rejected by hand during previous iterations of the loop (cf. 3.3.).

The third part of this step deals with prefixes. Indeed, decomposing the canonical form of a lemma in a prefix and a derivational basis can be very useful for a lot of tasks, including lexicon acquisition itself: a possible lemma is much more likely to be correct if it can be analyzed as a derivation of a known lemma by the adjunction of a standard prefix (this adjunction has usually to respect specific constraints[2]). Therefore, we try to decompose the canonical form of each generated lemma into one of predefined prefixes (the list depending on the part-of-speech) and the canonical form of an already generated lemma, with the additional constraint that the inflectional category has to be the same for both lemmas. If a prefix is recognized, two separate lemmas are generated, the first with a raw canonical form, and the second one representing explicitly its canonical form as the concatenation of a prefix and a derivational basis.

At the end of this first step, we obtain a list of form-lemma pairs, associating several possible lemmas to each form.

3.2. Ranking of lemmas

Lexical acquisition techniques must be able to extract not only very common lemmas, but also poorly attested lemmas, without too much over-generation (noise). This is achieved by ranking both lemmas and form-lemma pairs that justify them.

Formally, given a lemma l generating a form f, let $P(l)$ be the probability that l is a proper lemma (i.e. that the lemma l is attested in the corpus by at least one of its forms) and $P(f \prec l)$ is the probability that this lemma explains a given attested form f in the corpus. Let $\mathcal{L}_f$ be the set of all lemmas generating the form f, and $\mathcal{F}_l^{att}$ the set of all attested forms generated by the lemma l. We introduce the number of occurrences of a form f in the corpus $occ(f)$, and the number of occurrences of a lemma l in the corpus, which is:

$$occ(l) = \sum_{f \in \mathcal{F}_l^{att}} P(f \prec l).occ(f).$$

We also introduce the probability $P_l(f)$ for a form of the lemma l to be the form f (depending on l's inflectional class), and the probability $P(f|l)$ for a given form f to be attested in the corpus if the lemma l is attested. The latter satisfies the following equation:

$$P(f|l) = 1 - \prod_{l \in \mathcal{L}_f} \left(1 - P_l(f)^{occ(l)}\right).$$

Given these definitions, our algorithm is based on the sensible assumption that the following equations are satisfied :

$$P(f \prec l) = \frac{P(l).P(f|l)}{\sum_{l' \in \mathcal{L}_f} P(l').P(f|l')}$$

$$P(l) = 1 - \prod_{f' \in \mathcal{F}_l^{att}} (1 - P(f' \prec l)).$$

The first equation is a Bayes formula, and the second expresses the fact that a lemma l is not attested in the corpus if none of its attested (but ambiguous) forms is recognized as generated by l.

To compute approximatively the probabilities of lemmas, we use these equations in an iterative fix-point computation to solve these equations.

[1] As explained in the following part, the lexicon learning process is incremental. Therefore, the partial lexicon learned after the n-th iteration of the process could be used to learn anew the tagger, then tag the corpus again in a more accurate way, and use this newly tagged corpus for the $n+1$-th iteration. We didn't experiment this process yet because we decided to first improve the acquisition of the lexicon in a more drastic (and complementary) way, by taking into account globally cross-parts-of-speech morphological derivations. This work, currently in progress, will be published later.

[2] For example, in French, the verbal prefix "dé-" can correspond to English prefixes "un-": "défaire" means "to undo" and is based on "faire", which means "to do". However, if the derivational basis begins with a vowel phoneme, this prefix is extended with an extra "s": "obéir" ("to obey") is related to "désobéir" ("to disobey"). By "specific constraints" we mean this kind of phenomena.

3.3. Hand validation

The first iteration of the computation described in the previous paragraph leads to a result which is good but has noise, mostly because of misspelled or mistagged words in the corpus[3]. They can have two different consequences:

- one — or a couple of — misspelled or mistagged word can justify a wrong lemma which has these words as only form (”wrong-form-based lemmas”), as for example ”démesurer”, based on the noun ”démesure” and the adjective ”démesuré”, both mistagged as verbs,
- one — or a couple of — misspelled or mistagged word can influence a true lemma by giving it a wrong inflectional category and/or by changing its spelling (”misguessed lemmas”), as for example ”reconstruir”, based on several forms of the correct lemma ”reconstruire” and on the misspelled word ”reconstruir”.

Mistakes can also be errors of the acquisition algorithm itself, introducing incorrect lemmas with a probability $P(l) = 1$ that are based only on correct forms, as for example ”rendormer” instead of ”rendormir”. However, in our experiments, this occurs very rarely (see 4.).

Thus, and because any good quality lexicon has to be hand-validated anyway, we then ask a native speaker to validate the n best-ranked lemmas (n being usually 500 in our case). Three judgments can be given to a lemma: it can be set as valid, it can be set as ”wrong-form-based” because generated only by wrong forms, or it can be set as ”misguessed” because it is wrong although generating some correct attested forms. We must point out the fact that this hand-validation is not very costly, since lemmas are already well ordered, and since the validation task is extremely simple. This cost has nothing to do with the cost of previous methods to develop a lexicon, basically based only on human lexicographers. Furthermore, we facilitate hand-validation by grading lemmas l such as $P(l) = 1$ by the number $occ(l)$ of occurrences of the lemma l if it has no prefix, and by $occ(l) + 10$ if it has a prefix[4].

The results of this hand-validation is used to create both above-mentioned filters:

1. newly rejected lemmas, are appended to the list of rejected lemmas,
2. all forms of new wrong-form-based lemmas, and forms only explainable by new misguessed lemmas, are appended to the list of rejected forms.

The whole three-steps loop is then reiterated, until best-ranked not-yet-validated lemmas have a probability close to zero.

[3]Misspelled words are a tough problem, because they are unavoidable in a corpus. Furthermore some spelling mistakes are systematic, thus giving rise to consistent sets of (misspelled) inflected forms, which are best interpreted as a well-attested misspelled lemma.

[4]This incrementation by 10 if is the result of an empirical statistical analysis (depending on the corpus we used), so that the function giving the probability of correctness depending on the number of occurrences is the same both for lemmas with and without prefixes.

4. Preliminary results

On our general corpus, the fix-point algorithm takes a few minutes to converge, thus allowing fast iteration of the whole loop. The necessary manual validation is greatly helped by the quality of the plausibility measures computed. The first application of our method was to acquire the verbal lexicon of a 25 million words general (journalistic) French corpus. After the first iteration of the loop (thus without any filtering), the 100th incorrect lemma had rank 3337, and 7500 lemmas have a probability $P(l) \geq 1$, out of which 5250 have an number of occurrences at least equal to 1. It is noticeable that almost all incorrect lemmas come from misspelled or mistagged words (respectively 19% and 76% of the 100 best-ranked incorrect lemmas), and not from bad guesses of the algorithm based on correct forms (5%). The following table gives the number of correct and incorrect lemmas depending on their number of occurrences[5] after the first iteration of the loop, given by the comparison with the final acquired lexicon described hereafter. Of course these intermediate results are not supposed to be perfect, and are given only as illustration. Indeed, better results after the first iteration could be easily achieved with stronger constraints on acceptable forms (e.g. a form could be rejected if it occurs only once, or if it is tagged as non-verb more that a given percentage of the time), but it would reduce the number of learnable lemmas. Here, we give priority to lower silence over lower noise, because hand-validation will decrease noise more and more during the iteration process.

After one iteration, lemmas l such as $P(l) > 0.97$		
$occ(l)$	Correct lemmas	Incorrect lemmas
≥ 20	2831	66
≥ 10	3403	164
≥ 5	3771	277
≥ 2	4293	674
> 1	4390	807
≥ 1	4761	1728

Table 1: First intermediate results, before any hand-validation to reject misspelled and mistagged words. Almost all incorrect lemmas are caused by misspelled and mistagged words (respectively about 19% and 76% after this first iteration).

After only a few hours of manual validation and loop iteration, a lexicon of nearly 5000 verbs was achieved[6], generating approximately 200,000 forms, out of which several hundreds are not in the standard reference (Bescherelle, 1990), even though attested in a non-specialized corpus. Around 1500 verbs of (Bescherelle, 1990) are not acquired, most likely because they are not attested in the corpus, as suggested by the following check. In order to evaluate our work, we tagged our corpus using a ver-

[5]For lemmas with prefix, number of occurrences is increased by 10. See previous footnote.

[6]This number is slightly higher than the previously given 4759 (see Table 1) because we hand-validated also words with a number of occurrences and/or a probability $P(l)$ lower than 1.

sion of `TreeTagger` that we trained with an other freely available lexicon (Association des Bibliophiles Universels, ABU) ; only 1.64% of words tagged as verbs are not derivable from our acquired lexicon, which can be compared with the mean error rate of the this tagger, 3.25%. A manual check shows that these 1.64% are indeed almost always mistakes of the tagger. This lexicon of French verbs is the first part of our general French lexicon *LeFFF* (Lexique des Formes Fléchies du Français — Lexicon of Inflected Forms of French), available at `http://www.lefff.net/` under an open-source license published on the web site.

Our preliminary results on adjectives gives us approximatively 10000 lemmas from the general corpus. We then applied our algorithm on a botanic corpus of 3 million words: after the first pass, out of the first 1000 ranked adjectival lemmas, 350 are unknown to the previously built general lexicon, and almost all of them are indeed correct specialized terms.

5. Conclusion

While preliminary, our results demonstrate the feasibility of mechanized acquisition of wide-coverage lexica from general or specialized corpora, with very limited human labor[7]. The resulting wide-coverage French lexicon is made progressively available under a free-software license (at `http://www.lefff.net/`). Further work could include the improvement of the current algorithm, the acquisition of specialized lexica (as started on botanics), and the acquisition of lexica for other dialects of French or other languages with a rich morphology. In fact, we are in the process of globalizing our approach to be able to take into account derivational morphology, thus validating a lemma not only by its attested forms but also by its derivatives, in order to acquire the lexicon in a global way. Such a method should facilitate even more the fast acquisition of large-coverage lexica. It could even bring interesting linguistic results on derivational morphology, including subcategorization information through derivation links, and basic lexical semantics through linguistic knowledge about derivational morphology.

6. References

Association des Bibliophiles Universels, ABU. Dictionnaire des mots communs. In *La Bibliothèque Universelle,* `http://abu.cnam.fr/DICO/mots-communs.html`. Conservatoire National des Arts et Métiers.

Bescherelle, L.-N., 1990. *La conjugaison — Dictionnaire de douze mille verbes*. Paris: Hatier.

Briscoe, T. and J. Carroll, 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Washington, DC.

Daille, B., 2000. Morphological rule induction for terminology acquisition. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbrucken, Germany.

Dendien, J. and J.-M. Pierrel, 2003. Le trésor de la langue française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues*, 44(2).

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Grevisse, M., 1993. *Le bon usage, grammaire française*. Paris: Duculot, 13th edition. Revised by A. Goosse.

Ide, Nancy and Jean Véronis, 1994. MULTEXT: Multilingual text tools and corpora. In *Proceedings of the 15th. International Conference on Computational Linguistics (COLING 94)*, volume I. Kyoto, Japan.

Schmid, H., 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*. Manchester, UK.

[7] The small amount of human input (filtering of forms and lemmas) can be memorized, thus making the acquisition process reproducible, and opening it to scrutiny, comparison, criticism and improvement.