Inline Contraction Decomposition

Language Independent POS Tagging in the CorpusWiki project

Maarten Janssen
IULA, Universitat Pompeu Fabra
Barcelona
Maarten.Janssen@upf.edu

1 Introduction

Assigning POS tags to words in a text is typically done of three steps: tokenization, morphological analysis, and disambiguation. The last two of those steps are frequently done in a statistically-driven, language-independent fashion: although rule-based disambiguation tools contain language specific rules, most statistical disambiguation tools are in fact language independent. Even morphological analysis is done in a language-independent manner in many off-the-shelf POS taggers, where letter sequences are used to determine the possible POS tags for unknown words.

That means that the smallest of the three steps in POS tagging, tokenization, is the only part that is truly language dependent in current taggers. For the construction of a truly language-independent POS tagging system, tokenization is therefore a significant obstacle. Tokenization involves several problems, including sentences splitting, the dissection of contractions, and the treatment of multiword expressions, including the recognition of named entities and dates. In this article I will show how the language-dependent contraction splitting rules can be avoided by making tokenization part of the tagging process itself. This method was developed for CorpusWiki, an online language-independent tagging environment, which is presented in the next section.

2 CorpusWiki

CorpusWiki (http://www.corpuswiki.org) is an online tool that aims to allow (non-computational) linguists to build POS tagged corpora for their language of choice. The system is in principle meant to allow for the creation of tagged corpora for languages for which no corpus data exist, and for which it would be very difficult to create tagged data by traditional means (although it has been used for large languages like Spanish and English as well). For large but less-resourced languages there are often corpus projects under way, in the case of Geogian there is for instance the corpus project by Paul Meurer (2011), as well as corpora without POS tags, such as the dialectal corpus by Beridze & Nadaraia (2010). But for smaller languages such as for instance Ossetian, Urum, or Laz corpus projects of any size are much less likely. Corpora for these languages without POS tags often exist, for these specific languages in the TITUS project (Gippert), but annotating such corpora involves a computational staff that is typically not available for such languages.

CorpusWiki attempts to make it much easier to create an annotated corpora without the need of involvement of computational linguistic staff, by having a user-friendly, language-independent interface in which the user only has to make linguistic judgements, and the computational machinery is taken care of automatically behind the screens. The system is designed for the construction of gold-standard style corpora of around 1 millions tokens that are manually verified, and all corpora are build in a collaborative fashion, in which all users can help in enlarging the corpus of a given language. For all languages, both the corpus and the tagger with parameter files are available for download.

In the CorpusWiki set-up, each text in the corpus is individually treated in three steps. First, the text is added to the system. Then the text is automatically assigned POS tags using an internal POS tagger, which is trained on all tagged texts already in the system. And finally, the errors made by the automatic tagger have to be corrected manually. Once the verification of the tags is complete, the tagger is retrained automatically. In this fashion, with

each new text, the accuracy of the tagger improves and the amount of tagging errors that have to be corrected goes down. The only text that is treated differently in this set-up is the very first text, since for the first text, there are no prior tagged data. The system uses a canonical fable as the first text for each language to make the initial manual tagging of the first text go as smoothly as possible.

To make the tagging process more user friendly, tags are not presented as position-based tags, but rather as separated features, with a pull-down presenting the possible values for each feature, such as "singular/plural" for number on nouns (in languages with a binary number system for nouns). Which are the possible values in a given language is defined by the user, who indicates which classes, features, and values are used in that language, by selecting them from a long list of possible grammatical features and values. By having pre-defined pull-downs, the user verifying the tagging errors does not have to study the tagset beforehand, but merely has to select the correct option from amongst a set of easily recognized options.

3 Contractions

A contraction in the computational sense is a single orthographic word that grammatically functions as two separate words. A good example is the French words *au*, which is the contraction of a preposition (*à* = at, in) and the masculine definite determiner. There are three common ways of treating such contractions: by assigning a single, complex tag to the word as in (1), or splitting the word into two separate words as in (2), potentially while keeping the information about the contracted words the individual parts, as in (3).

(1) au PREP+DET

(2) à PREP
le DET

(3) à_au PREP
le_au DET

In many phonologically transcribed corpora, option (1) is preferred, since the contraction forms a single phonological word, which makes it hard to represent phonological transcription in (2) and (3). On the other hand, in (1) it is difficult to represent the lemmatized form of the two separate word, which is one of the reasons why (2) is more common in written corpora.

In the solution in (2), the contraction is simply treated as if it were two separated words. The separation of contractions is done in the tokenization process, so already at the level of morphological analysis, there are no longer any contractions in the text. As said before, in order to establish this, it is necessary to write language-specific rules that tell which strings in the language should be separated, which is not a feasible option for a system like CorpusWiki.

Instead, CorpusWiki pushes the separation of contractions further down the pipe, and only deals with word separation as part of the morphological analysis, hence getting rid of tokenization as a distinct step altogether. That is to say, the initial process only separates words simply on spaces (and punctuation marks), and only in a later stage do some of the space-delimited tokens that are in fact more than one word get subdivided into several tokens.

Not all contractions are alike: which words are written together differs greatly from language to language. In order to correctly deal with tokenization, CorpusWiki distinguishes between four different types of contractions:

1. *Lexical contrations*. Lexical contractions are individual combinations of words written together, such as for instance the case of *au* in French. They are not always transparent (there is no recognizable orthographic trace of the determiner *le* in *au*), and they are not always separable (*à* in Portuguese is a contraction of the preposition *a* and the determiner *a*, but there are no separate parts that can be assigned the role of PREP and DET). They do typically have a lexicalized form that is listed in the dictionary (*aux* is the plural form of *au*, which has a lexical entry in most French dictionaries).

2. *Orthographic contractions*. Orthographic contractions are productive and open-ended. For instance, whenever the French determiner *le* is followed by a word starting with a vowel, it is written together with that

word: *le* + *eau* = *l'eau* (the water). The parts of an orthographic contraction are clearly recognizable, and in the vaste majority of cases, there is one non-modified base word (*eau*), and another (trunctated) word adjoined to it (*l'*). Orthographic contraction do not have a citation form, and are never truly seen as one word.

3. *Clitics*. Clitics are in most respects like orthographic contractions, except that linguistically, they are viewed as something between an affix and a separate word. Most contractions do not have a lexicalized form, but there are cases in which the clitized word is typically said to have a lexical form. For instance, reflexive verbs with a clitic in Romance languages are typically represented with a cliticized citation form: in the Catalan sentence *vols rentar-te?* (do you want to wash yourself), the verb *rentar-te* is said to be a form of the pronominal verb *rentar-se*.

4. *Agglutinates*. Agglutinative morphemes are part of the inflectional morphology, and as such, should not be seen as contractions. However, from a structural perspective, some agglutinating affixes should be seen as contractions nevertheless, mostly class-changing affixes. Take the Turkish word *odamdayım* (I am in my room). Taken as a whole, this is a verb form in the 1st person singular indicative. But it does not have a lemmatized form as a verb, only a (lemmatized) root form: *oda* (room), which itself is not a verb but a noun. The verbal part of the word (*-yım*) behaves in several respects as separated from the nominal base. For instance, when an adjective, such as *kücük* (small) is placed in front of the word, will modify only the noun, and not the word as a whole. Therefore, the part *-yım* is treated as a contracted part as well, in concordance with the treatment by for instance Bisazza & Federico (2009).

Although there are differences in treatment between all these types of "contractions" in CorpusWiki, for the purpose of separation only the contrast between (1) and (2)-(4) is of importance. Lexical contractions are simply stored in the corpus as contractions, and when a lexical contraction is found in a new text, it will only be separated if it already was used as a lexical contraction in the training corpus. The productive types of contractions (2-4), on the other hand, are treated in the morphological analysis step.

4 Productive contractions

During the morphological analysis, productive contractions get separated in the following way: any word that can potentially be a contracted word is assigned a contracted structure as (one of) its morphological analyses, together with the lexical likelihood of that analysis. The potential contractions, like the rest of the morphological analysis in CorpusWiki, are established on the basis of the training corpus.

To establish this, contractions in the corpus are subdivided into sub-tokens, while keeping the space-delimited token as well. An example of the structure for the French contraction *l'eau* (the water) can be found in figure 1, presented in the internal XML-based format of the system. In this example, the space-delimited token is marked as a (orthographic) contraction, and the two individual tokens *l'* and *eau* are marked as a determiner and a common noun, respectively, and provided with a lemma as well as a part of the contracted form. Furthermore, the first part (l') is marked as a contracted/truncated part, and the second as the unmodified base. Suppressed in figure 1 are the additional morphosyntactic feature/value pairs, indicating that *l'* is the masculine singular form of the definite article.

```
<tok pos="CONTR" & type="orthographic">
     l'eau
     <dtok pos=DET lemma="le" dtype="contracted" form="l'"/>
     <dtok pos="SUB" lemma="eau" dtype="base" form="eau"/>
</tok>
```

Figure 1: Figure 1. CorpusWiki representation for *l'eau*

The subdivision in the training corpus is done by the annotator in two steps: first, the annotator has to indicate that the word is a contraction, and which of the four types of contractions it is, and then each part of the contraction

is treated like any other word with pull-downs for feature-value pairs. And as with normal words, the percentage of cases in which the tagger will already correctly analyze the contraction correctly will grow as the corpus gets bigger.

When building the parameter files from the training corpus, the system will collect all contracted forms, and store them in a file (contraction lexicon), with an indication about their frequency, and whether they appear before or after the base form. So from figure 1, the system will add a record for a prefixing contracted form *l'* to the contraction lexicon.

When parsing a new text, this contraction lexicon is then used to parse (potential) contractions in the following way. When encountering the word *l'eau*, the system will system mark it as it was marked in the previous text: as an orthographic contraction of the determiner *l'* and the noun *eau*. But when dealing with a word that has not appeared in the training corpus, say *l'administrateur* (the administrator), the system will attempt to treat it as a contraction, since it starts with a string that is in the contraction lexicon.

The probability of the analysis of *l'administrateur* as a contraction is calculated on the basis of all the words in the training corpus that start with *l'*, and the likelihood of *l'administrateur* being a contraction depends on the percentage of those words that is in fact a contraction. Since all words in the training corpus starting with *l'* will be contractions (since the apostrophe in French is an explicit truncation marker), the tagger will mark all new occurrences starting with *l'* as contractions as well. Notice that this is not based on the presence of the apostrophe itself, since there are many other languages in which the apostrophe does not mark contractions.

In cases (languages) where there is no explicit marking of the truncation, the system will decide in the disambiguation process whether the word is more likely to be a contraction or a simple word. So for instance, the Spanish word *hazte* (do/make yourself) is a form of the verb *hacer* (to do) with the clitic *the* (you) attached to it, without an explicit boundary between the two. Since there are many other words in Spanish ending on *-te* that are not cliticized words, the system will keep both options, with their respective lexical likelihood score, for disambiguation in context. Hence, as OOV items, both *parate* (stop yourself) and *karate* (karate) will be fed to the disambiguation step as either cliticized verbs or substantives.

After splitting off the contracted part (clitic, truncated form, or agglutinate), the system will continue with the morphological analysis of the remainder of the word. So when parsing the word *l'eau*, the system will first split the word in two parts, end then determine the potential POS tags for both parts the same way it does for any word. In this particular case, it will find that *l'* can be a masculine or feminine definite article, and *eau* can be only a feminine singular common noun, and determine the most likely of these in the disambiguation process.

Since this process is recursive, it can deal with words with multiple contracted parts without much problem. So when dealing with a word with two clitics, like the form *haztelo* (*haz* (do) + *the* (you) + *lo* (it)) in Spanish, it will first split off the indefinite pronoun *lo* and after that the personal pronoun *te*. And since contraction parts can appear on either side, there is no problem with words that have contracted parts on both sides either, such as the Catalan *d'anar-hi* (to go there), where the base form *anar* (to go) has an orthographic contraction to the left, and a clitic to the right.

Because the separation of contracted forms in this method is driven completely by examples provided in the training corpus, the process is dealt with in a statistically-driven, language-independent fashion. The overall result is in (almost) all cases the same in the end as a method in which language-specific tokenization rules are written, but works without the need to write explicit rules. This means that it is a tokenization method that is ideally suited for a language-independent system like CorpusWiki.

5 References

Marina Beridze and David Nadaraia. 2010. The corpus of Georgian dialects. In *Proceedings of fifth International Conference*, Slovakia.

A. Bisazza and M. Federico. 2009. Morphological pre-processing for Turkish to English statistical machine translation. In *IWSLT 2009 - International Workshop on Spoken Language Translation*.

Paul Meurer. 2011. Constructing an annotated corpus for Georgian. paper presented at the Ninth Tbilisi Symposium, Kutaisi, 2011.