MaxMatch Tokenizer Implementation

Lorenzo Tosi

November 3, 2018

Contents

1 Introduction

For this Practical 2 segmenters were used: *Pragmatic Segmenter* and *NLTK*. The text to analyze was in Emilian-Romagnol, a continuum of dialects with similar traits and grammar spoken in Emilia-Romagna, a region of Northern Italy. Both segmenters, given the small diffusion of the language, are not supporting it natively.
The segmenters were set first in English and then in Italian, the results show no differences between the two options.

2 How the Analysis Was Run

Sentences were extracted through *WikiExtractor* and then filtered using this bash code:

```
sed -e '/.\{450\}/!d' < wiki.txt | sed 50q > wikifiltered.txt
```

I chose to use only paragraphs longer than 450 characters as the Emilian-Romagnol Wikipedia is composed of many single-sentence articles, that are not very useful for our case study.
The *Pragmatic Segmenter* program was created as indicated in the Practical guidelines.
Here is the code for the *NLTK*program adapted for choosing different languages.

```
import nltk.data
tokenizer = nltk.data.load("tokenizers/punkt/italian.pickle")
with open("wikifiltered.txt", "r", encoding="utf-8") as f:
        l = f.read()
        token_list = tokenizer.tokenize(l)
        print(len(token_list))
with open("wikisegmented_nltk.txt", "w+", encoding="utf-8") as s
        for x in token_list:
                s.write(x + "\n")
```

3 Comparison

Programs testing was done through a Python script called *comparison_test.py* that I attach in the package.

Test file The file with correct hand-made segmentation has 109 sentences.

Pragmatic Segmenter 103 sentences, 96 out of 109 correspond to the test file.
The segmenters struggles when word starting with a ' appear, as in both English and Modern Italian words do not start with an apostrophe. For example in this sentence:

> 1 L'é alzêr a Burèt e Gualtēr dóve a 's mantînen i sòun ö e ü in ûş in Lumbardìa. A rişûlta pèiş a Guastâla, Lusêra e Rezōl dóve l' é parlê, cun pôchi sfumadûri tra i paèiş, al guastalèiş, sòt gróp dal dialèt mantvân e un bèl pô divêrs da l'arzân.

Seems that the segmenter considers the part from *"a 's mantînen"* to *"l' é parlê"* as part of a single quoted speech sentence, as the apostrophe resembles the position of a quotation mark (space, apostrophe, letter — letter, apostrophe, space), and this segmenter does not segment reported speech. Another example of it is in this sentence:

> 1 ""...l'ôpra dl'artésta, se êrt l'é, l'an sèrov mia per dîr un quél, a fêr dal dichiarasiòun, mó per tirêr fôra da ciaschidûn ed nuèter còl che agh'om dèinter, biânch o nîgher ch'al sia, scûr o aleghēr..."" e che: ""...fôrsi dèinter int 'na figûra, e mìa sôl int 'na figûra, an n'é mìa impurtânt còl che gh'é, còl ch'la fà vèder. Despès l'é pió impurtânt còl ch'agh mânca, lasând a nuèter la posibilitê ed serchêrel, sfumadûra o òm ch'al sia.""

where dots inside the quotation marks are not segmented.
The segmenter also do not segment the abbreviation D.O.C. placed on sentence boundary:

> 1 Al vèin 'd la cusèina arzâna, che bèin a's acumpâgna al sô bâgni l'é al Lambrósch Arzân, ròs ch'al spóma naturêl, incô protèt da un mêrch

D.O.C. Al Biânch de Scandiân, ed góst dôls e mèz sèch, l'é fât ind al tîpo ch'al spóma e al spumânt.

Another issue is segmenting Ecc.. (Etc. in English) as the abbreviation contains two dots:

1 A tèvla a catòm al carèl dal lès cun sampèt, cudghîn, sampòun, ecc.
2 ., in sèma al bânch di salumêr la vaschèta di grasōl, salâm e persót.

NLTK 110 sentences, 108 out of 109 correspond to the test file.
NLTK does a great job at segmenting the chosen Wikipedia dump. NLTK is even able to recognize abbreviations as Ecc... and D.O.C. (Denominazione di Origine Controllata) placed on sentence boundary and segment correctly:

1 Al vèin 'd la cusèina arzâna, che bèin a's acumpâgna al sô bâgni l'é al Lambrósch Arzân, ròs ch'al spóma naturêl, incô protèt da un mêrch D.O.C.
2 Al Biânch de Scandiân, ed góst dôls e mèz sèch, l'é fât ind al tîpo ch'al spóma e al spumânt.

The only problem is not recognizing that the last dot of T.A.V. (Treno ad Alta Velocità) is not a sentence boundary, even if enclosed in round brackets:

1 In pió a s' é drê fêr la nōva stasiòun pr' al trêno a êlta velocitê (T.A.V.)
2 ch' la gh' à al nòm de Stasiòun Mèdio Padâna (...)