Pràctica1: Etiquetatge gramatical basat en unigrames

1 Introducció

En lingüística de corpus, l'**etiquetatge gramatical** —que en anglès s'anomena *part-of-speech tagging*, *grammatical tagging* o *word-category disambiguation*— és el procés en el qual es marca una paraula en un text amb la seva categoria gramatical corresponent. Mitjançant un *etiquetador* o *anotador*, donada una seqüència de paraules, s'assigna una seqüència d'etiquetes gramaticals (*tags*):

```
El nen menja una poma + [ETIQUETADOR] = DET N V DET N
```

La principal dificultat del *tagging* és l'ambigüitat. La correspondència paraula-tag no és sempre unívoca, de manera que una mateixa paraula pot tenir etiquetes diferents en frases diferents. Així doncs, la categoria gramatical d'una paraula en una frase depèn del seu context, tal com es pot veure en el següent exemple:

Ha pesat/V tres kilos.
El professor és un pesat/N.
Un discurs pesat/ADJ.

1.1 Automatització del procés

- Opció 1: Regles
  - Obtenir un diccionari
  - Escriure les regles corresponents

    Exemple: Si $w_i =$ "*metge*" i $t_{i-1} = DET$, aleshores $t_i = N$.
- Opció 2: Estadística
  - Utilitzar *machine learning* (aprenentatge automàtic).

1.2 Tagging basat en aprenentatge automàtic

Corpus En primer lloc, si es tracta d'un procés supervisat, necessitem un corpus preliminar anotat manualment, que utilitzarem com a text d'entrenament. En lingüística computacional s'acostuma a anomenar *corpus anotat*.

Objectiu Extreure coneixement d'aquestes dades per poder repetir l'operació d'etiquetatge amb altres textos.

Fases

- **Entrenament**: A partir d'un corpus anotat, extraiem un algorisme que ens generarà un **model del llenguatge**.
- **Etiquetatge**: Utilitzarem el model del llenguatge generat per etiquetar qualsevol text nou.
- **Avaluació**: Compararem el text etiquetat automàticament amb un text etiquetat manualment per veure'n el percentatge de correcció.

Anotació estadística

- Donada una seqüència de paraules $W = (w_1, ..., w_n)$ i una seqüència de tags $T = (t_1, ..., t_n)$, definim la probabilitat que a W li correspongui T com a: $P(T|W)$
- Considerem que la millor seqüència serà la que maximitzi l'expressió anterior: $\hat{T} = argmax_T P(T|W)$
- La realitat (aplicant la regla de la cadena):
  $P(T|W) = \Pi_{i=1}^{N} P(t_i|w_N, ..., w_1, t_{i-1}, ..., t_1)$

2 Mètodes estadístics. Implementació pràctica

2.1 Unigrames

L'etiquetatge de text basat en **unigrames** és el mètode estadístic més simple. A l'hora d'etiquetar una paraula en una frase la considerarem de forma independent i li assignarem l'etiqueta més probable d'acord amb un corpus d'entrenament (corpus anotat).

És a dir, si la paraula *cantat* apareix en un corpus d'entrenament 768 vegades, de les quals 443 actua com a adjectiu ('Adj') i 325 com a verb ('V'), sempre li assignarem a *cantat* l'etiqueta 'Adj'.

Així doncs, a la pràctica farem:

$$P(T|W) \approx \Pi_{i=1}^{N} P(t_i|w_i) \tag{1}$$

Estimació de les probabilitats

Probabilitat lèxica La probabilitat que una paraula aparegui a la posició i depèn només del tag a la posició i.

1. $C(t_i)$: Número d'ocurrències del tag t_i.
2. $C(w_i \wedge t_i)$: Número de vegades que la paraula w_i apareix etiquetada com a t_i.

aleshores $P(w_i|t_i)$ s'estima com:

$$P(w_i|t_i) = \frac{C(w_i \wedge t_i)}{C(t_i)}$$

3 Realització de la pràctica

Per grups de 2 o 3.

Etiquetatge basat en unigrames

En aquesta primera pràctica etiquetarem un text utilitzant el mètode estadÃstic més simple possible. Per a això, en el moment d'etiquetar una paraula en una frase la considerarem de forma independent i li assignarem l'etiqueta més probable en funció d'un corpus d'entrenament.

Primer pas: Generació del Model del Llenguatge El primer pas consisteix a escriure un programa que, donat un corpus anotat (en el nostre cas, el fitxer `corpus.txt`), escrigui en un fitxer la informació dels tags i les paraules trobades. El fitxer s'anomenarà "lexic.txt' i cada línia ha de contenir la informació (paraula, tag, ocurrències) separada per tabuladors.

```
cantado    Adj      443
cantado    V        325
cantados   Adj      13
...
```

Segon pas: Etiquetatge d'un corpus utilitzant un Model del Llenguatge. Per a completar aquesta part de la pràctica s'ha d'implementar un programa que utilitzant el Model del Llenguatge generat a l'apartat anterior etiqueti els fitxers `test\_1.txt` i `test\_2.txt`.

Tercer pas: Avaluació dels resultats El tercer pas consisteix a avaluar els resultats del programa. Per a això cal escriure un programa que compari els resultats del nostre etiquetatge amb l'etiquetatge *correcte*. Aquests resultats correctes es troben als fitxers `gold\_standard\_1.txt` i `gold\_standard\_2.txt`. Aquests fitxers són la referència; si el programa etiquetés totes les paraules del text correctament, els resultats coincidirien 100% amb el *gold standard*. El programa ha de comparar els dos resultats i indicar el percentatge de correcció. **L'informe i el codi s'han de lliurar el dia 12-03-2017 a les 23:55h**.