Manual para participantes

From Tweet-norm

El proceso de etiquetado se puede resumir en los siguientes puntos:

PREPROCESO

1) Los tweets se procesan con *Freeling* (http://nlp.lsi.upc.edu/freeling/).

```
  analyze -f es.cfg --flush --ftok es-twit-tok.dat --usr --fmap es-twit-map.dat  --outf morfo --noprob --noloc
```

donde los ficheros *es-twit*.dat* están en el SVN: http://devel.cpl.upc.edu/freeling/svn/trunk/src/main/twitter/

2) Las palabras sin análisis se marcan como OOV.

3) Las palabras OOV son las que se anotan, se distribuyen y se evalúan. Por lo tanto las palabras real-word-errors no son tenidas en cuenta en ningún momento.

4) En la anotación se cambia la marca OOV por una de las siguientes: VARIATION, CORRECT y NoES (no español). En caso de VARIATION se le asigna su correspondiente normalización.

5) VARIATION, CORRECT y NoES se exportan como tipo 0, 1 y 2.

REGLAS DE ANOTACIÓN

CASO 1: Palabra incluida en la RAE

Regla 1: en todo caso se anotará como correcta sin modificación alguna, aunque por su contexto se dedujera que dicha palabra no es la adecuada.

CASO 2: Palabra no incluida en la RAE con categoría de nombre propio

Regla 2.1: Si es un acrónimo [*formado por todas las letras requeridas según el contexto*], originalmente compuesto con alguna minúscula, tanto la forma originaria como su forma totalmente en mayusculas serán etiquetadas como correctas sin ninguna modificación.

```
 (p.e., CoNLL y CONLL)
```

Regla 2.2: Si es un acrónimo [*formado por todas las letras requeridas según el contexto*] y originalmente compuesto con todas las letras en mayúscula, será etiquetado como correcto sin modificación alguna.

```
 (p.e., IBM e I.B.M.)
```

Regla 2.3: Si no es un acrónimo, está formado por las letras requeridas y su inicial está en mayúsculas e incorpora los acentos requeridos, será etiquetada como correcta, ya sea un nombre propio en diminutivo, un apodo u otra forma

alternativa de la entidad.

```
(p.e., Tony, Anita, Yoyas)
```

Regla 2.4: Si se expresa con alguna falta de ortografía o con alguna alteración no aceptada por las reglas 2.1 a 2.3 , se anotará como variante y se especificará su forma correcta, según se define con dichas reglas.

```
(p.e., sanchez -> Sánchez, tamagochi -> Tamagotchi, abc -> ABC, a.B.c. -> A.B.C.,
[*CONL -> CONLL*])
```

CASO 3: Palabra no incluida en la RAE con otra categoría diferente a nombre propio

Regla 3.1: Si es un neologismo o estranjerismo compuesto correctamente según reglas de buena formación se etiquetará como correcta sin ninguna modificación.

```
(p.e., mouriñistas, retuitear, retweetear)
```

Regla 3.3: Si es un diminutivo o superlativo compuesto correctamente según reglas de buena formación se etiquetará como correcta sin ninguna modificación.

```
(p.e., supergrande)
```

Regla 3.4: Si se expresa con alguna falta ortográfica o alteración (repetición, eliminación, permutación de letras, etc), se etiquetará como variante y se especificará su forma correcta.

```
(p.e., horrooorr -> horror, hacia-> hacía)
```

Regla 3.5: Si es una abreviatura o un acortamiento se etiquetará como variante, especificando su forma correcta.

```
(p.e., admin -> administración, sr -> señor)
```

Regla 3.6: Si es una onomatopeya con alguna alteración (normalmente repetición de letras) de una o varias forma existente según la RAE, entonces se etiquetará como variante de todas esas formas. Si no existe en la RAE se anotará como correcta.

```
(p.e., aaaahhh -> ah, jajajajas -> ja, iiiii -> uy+ay)
```

Regla 3.7: Si es una concatenación de palabras, entonces se etiquetará como variante y se especificará la secuencia correcta de palabras correctas.

Regla 3.8: Si es una palabra de otro idioma, se etiquetará como NoEs

```
(p.e., parking)
```

Regla 3.9: Si es una emoticon, se etiquetará como NoEs

Regla 3.10: Si es una cadena de palabras de otro idioma, se etiquetará entre NoEsBeg ... NoEsEnd