LINGI2263 : Computational Linguistics

Group 2 : Project 2

Authors:
Crochelet Martin (2236-10-00)
Baugnies Benjamin (6020-10-00)

Professor:
Pierre Dupont
Cédric Fairon

2013 - 2014

1 Tags and words statistics

The first part of the assignment consisted of creating a Lexicon based on the training corpus, and extracting some general information about the words and tags it contains. Our lexicon is limited to the 5000 most common words. Other words are replaced by the `<UNK>` token in both the training and test files. Among these words, we extracted the 10 most common words. They are summarized along with their frequency in table 1.

Rank	Word	# occurences
1	THE	59466
2	,	49639
3	.	41859
4	OF	31067
5	AND	24709
6	TO	22190
7	A	19718
8	IN	18230
9	THAT	8974
10	IS	8623

Table 1: Most common words

While parsing the texts, we found 187 different tags. The 10 most and least common of these can be found below 2.

	Most Common			Least Common	
Rank	Tag	# occurences	Rank	Tag	# occurences
1	NN	143023	-1	JJR+CS	1
2	IN	104385	-2	JJ$	1
3	AT	84266	-3	NN+HVD	1
4	JJ	58365	-4	RBR+CS	1
5	.	51997	-5	IN+IN	1
6	,	49641	-6	NR+MD	1
7	NNS	49412	-7	IN+NP	1
8	NP	32992	-8	WRB+BER	1
9	CC	32496	-9	WRB+MD	1
10	RB	31147	-10	NP+MD	1

Table 2: Most and least common tags

We can see that the most rarely seen tags are almost all compound tags. Moreover, the list of least used tags is actually not very precise since a lot more than 10 tags are used only once. The presented list here is just a chosen subset of those tags but we could have chosen other tags.

Finally, we have a count of the number of tokens and segments in each file (both training and test). We added the number of types as we will talk about it later for the uniquely tagged words.

	Training	Test
Tokens	987341	173524
Segments	48461	8552
Types	52016	6550

Table 3: Token & segment count

2 Baseline Tagger

For this part of the assignement, we were to tag all the words of the test file using only the most common tag for each word. To do this, we first create and copy of the "lexiconized" test file (with '¡UNK¿' for words that are not in the top 5000) in which we removed all the existing tags. We then, for each word in the lexicon, choose a single best tag in terms of number of occurences for that particular word.

For example, we have a entry for 'THROUGH' in our lexicon with two possible tags: 'RP' with 56 occurences and 'IN' with 781. The 'IN' tag is chosen, and therefore, all occurences of 'THROUGH' in the test file will be tagged as such.

Once the while document has been tagged using the same conventions as in the original, we compare the tags to those in the lexiconized document. Out of 173524 tokens, this baseline tagger tagged 162905 correctly for 10619 errors, given us an error rate of 6.12%. This is slightly higher than we expected. Indeed, we had found that in the training data, 41125 of the 52016 types were uniquely tagged. This represented 392544 out of the 987341 tokens, or almost 40%. Due to this large number of uniquely tagged tokens, the average word has only 1.1243 tags. This means that choosing randomly among the observed tags (as opposed to the best tag like we have done) would yield an accuracy of 88.94%.

The average error per tag calculated by

$$\frac{\sum_{tags} errors(tag)/occurences(tag)}{\#tags}$$

was of 18.47%. This is much higher than the global error rate, and therefore indicates that rarely used tags are more prone to error.

Finally, we can have a look at some tags in particular. We first take the 'JJS' tag. This tag is used 44 times in the test file, but the 'NN' tag was chosen instead 19 times (43.18%), giving it an accuracy of 56.82%. On the other hand, the 'NP' tag was used 2972 times. While more distinct tags were given by error (the two most common being 'JJ' 81 times or 2.73% and 'NN' 112 times or 3.77%), the tag had a 91.42% accuracy. This seems to confirm what was found in the previous paragraph.

3 HMM Tagger