FNLP Tutorial 1

1 Ambiguities

Ambiguities are pervasive in natural language, but often go unnoticed when we use language because humans are so good at resolving them. In this exercise, we want you to find ambiguities in the example sentences and attempt to articulate a paraphrase that as much as possible removes the ambiguity (similar to section 1.2 of J&M, 2nd edition). Categorise the different ambiguities you observe: e.g., word sense ambiguity, structural ambiguity, phonetic ambiguity and so on.

1. At the bank, Mary noticed her sister.
2. Every student wants to win the first prize in a programming competition with a robot.

2 Corpora and annotation

In this exercise, we want you to get some insights into the challenges that humans and machines face when it comes to annotation. Consider the following corpus:

1. Paris Hilton stayed at the Hilton in Paris.
2. Donald Fucking Trump.
3. James Clerk Maxwell was educated at Edinburgh and Cambridge.
4. Tom works for the Dumfries & Galloway Standard.

1. Annotate the above utterances with named entities. For our purposes, a named entity is a single word or multiple words that refer to a person (`PER`), location (`LOC`) or organisation (`ORG`). Are there cases that you found difficult? Which cases do you think are difficult for an automated system? And why?
2. Annotations in NLP must be mathematical objects in order to be machine-readable. How would you formalise the annotation of a named entity (e.g. using tuples, sets, lists, strings and natural numbers)? Provide formalised examples of two of the sentences. Are there any examples where you cannot unambiguously represent your annotations?

3 Preprocessing Twitter data

Read through the ten messages from the social media platform Twitter, below.[1]

Mini Twitter corpus

1. @fakeusername cool!I always lookd for one but only found Haighs and Koko Black -not that there's anything wrong with that ;)
2. sorry might not b back 4 a while i have alot coming up ,pantomine,option choices,holiday,footie play offs AND all my ruin your halfterm work
3. Let me sneak out to kitchen. I'm hella hungry. Brb!
4. @fakeusername yo im bored imma bout to call ya ass on skype
5. http://fakeurl.com - BITCHES DONT KNOW BOUT MY MILLENNIUM FALCON #fakehashtag
6. RT @fakeusername: RT @fakeusername: RT @fakeusername: Pleaseeeee, i want to be taller :(http://fakeurl.com
7. OH SNAP. my 6 year old cousin snores SO loud.
8. the an-noy-ing cramps is getting my way! I wanna kill eeuuu!
9. Daughter's been *kil-ling* it on video chat with Grandma. I'd put her on that ChatRoulette thing if that weren't, like, awful parenting.
10. Thr R times, tht I test yr faith,'til U think U might surrender. Baby Im, Im not ashamed 2 say, tht my hopes wr (cont) http://fakeurl.com

1. Imagine you want to POS-tag these sentences with a computational model trained using white-space tokenised articles from news papers. Rewrite the ten tweets in the format you think is best given this application.
2. In practice, preprocessing need not only be effective, but also cheap in terms of resources that are involved. Provide five rules for string manipulation that, in your opinion, will have the greatest impact. You can simply phrase them in text, or use regular expressions if you are familiar with them (see J&M, chapter 2, 3rd edition).

[1] Tweets are adapted from the SentiStrength corpus, that is free for academic use.