Presentation Summary:

- Have presented a statistical approach that uses HMM to do POS tagging of Arabic text.
- Have analyzed the Arabic language quite systematically and have come up with a good tag set of 55 tags.
- Have then used Buckwalter's stemmer to stem Arabic corpus and we manually corrected any tagging errors.
- Designed and built an HMM-based model of Arabic POS tags.
- One of the greatest advantages of having a trainable POS tagger is that it will speed up the process of tagging huge corpora.

References:

Fatma Al Shamsi, Ahmed Guessoum. "A Hidden Markov Model –Based POS Tagger for Arabic". JADT 2006 : 8es Journées internationales d'Analyse statistique des Données Textuelles.

Obstacles:

- It was hard to manage the flow of the points discussed in the paper to be presented in the class due to condensed information in this paper.
- New concepts had to be understood and researched to provide a point of comparison such as SVM and LCD.
- I was unable to fetch the full HMM Model and the statistical table to have a full picture of what is the full picture.

Lessons Learned and Recommendation :

- That some new technologies seem simple but in fact the work behind it is huge.
- That fact that you can tag a corpus with a performance higher than 90% and represent state of the art technology.
- Using this tagger in variety of in information retrieval, machine translation project transfer module, etc.
- This paper provides a better alternative than to the ones described in the lecture.
- Finally, the tag set shows how you can use a polymorphic analysis to minimize the tag sets, such as the tag SUFF_SUBJ_ALL.

Question/Answers:

Q- Define POS tagging?
A- It is the process of assigning a part-of-speech tag such as noun, verb, pronoun, preposition, adverb, adjective or other tags to each word in a sentence.

Q- What are the two approaches for POS tagging?
A- Rule based and trained ones. Rule based→ a knowledge base of rules is developed by linguists to define precisely how and where to assign the various POS tags. Trainable based→ statistical language models are built, refined and used to POS tag the input text automatically.

Q- What is the difference between the lexical and contextual ambiguity?

Lexical ambiguity occurs when the same word with the same constitutes has different meaning and is solved by using POS, and for example, in Arabic we use diacritics to resolve this matter. Contextual ambiguity comes form the different meanings of a sentence.