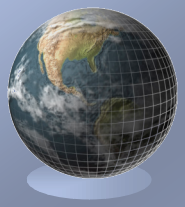
My Grad School Experience



By
Shereen Khoja

Introduction

- I will talk about:
  - Research for my M.Sc.
  - Research for my Ph.D.
- This will involve talking about:
  - Stemming
  - Tagging
  - The Arabic Language

M.Sc

- I started my M.Sc at the University of Essex in 1997

University of Essex

the campus

University of Essex

- The University of Essex is located in Colchester in the south east of England

Colchester

- Colchester is Britain's oldest recorded city

Colchester

The University of Essex

- The university has 9,100 students, 25% in the graduate programs
- The computer science department has 34 faculty members
- There were 120 students on the M.Sc. program

M.Sc. Courses

- I completed 7 courses while I was at the University of Essex
  - Computer Networks
  - Computer Vision
  - Expert Systems
  - Distributed AI & Artificial Life
  - Machine Learning
  - Neural Networks
  - Natural Language Processing

February 7, 2005

M.Sc Research

- I decided to do my research in the area of Natural Language Processing
- Natural Language Processing is a broad area of AI that focuses on how computers process language
- NLP has many sub-areas such as:
  - Computational Linguistics
  - Speech Synthesis
  - Speech Recognition
  - Information Retrieval

February 7, 2005

M.Sc Research

- My M.Sc research was conducted under the supervision of Professor Anne De Roeck
- She was already supervising a Ph.D student who was working on an Arabic language system
- It was decided that I would work on an Arabic language stemmer

February 7, 2005

Stemming

- What is stemming?
  - Stemming is the process of removing a words' prefixes and suffixes to extract the root or stem
  - Computers
  - Compute
  - Computing

February 7, 2005

Uses of Stemming

- Compression
- Text Searching
- Spell Checking
- Text Analysis

February 7, 2005

Arabic Language

- Arabic is an old language
- The language hasn't changed much in 1400 years
- Arabic is written from right to left

February 7, 2005

Arabic Language

- Arabic is a cursive language
- The shape of the letters change depending on whether they are at the beginning, the middle or the end of the word

Arabic Language

- 28 consonants in Arabic
- 3 of these are used as long vowels
- A number of short vowels or diacritics
  - drs
  - darasa
  - durisa
  - dars

The Arabic Language

- Arabic is a Semitic language, so words are built up from, and can be analysed down to roots following fixed patterns
- Patterns add prefixes, suffixes and infixes to the roots.
- Examples of words following the pattern Ma12oo3:
  - ktb maktoob
  - drs madroos

The Arabic Stemmer

- I developed an Arabic stemmer
- The stemmer used language rules
- These rules are based on the Arabic grammar, which hasn't changed for 1,400 years
- The stemmer was developed in Visual C++

Difficulties

- Words that do not have roots
- Root letters that are deleted
- Root letters that change

The Arabic Stemmer

- I released the stemmer under the GNU public licence
- The stemmer has been used by the following:
  - The University of Massachusetts
  - MitoSystems

Ph.D.

- I began my Ph.D. research at Lancaster University in 1998

February 7, 2005

Lancaster University

- Lancaster University is located in the city of Lancaster in the north west of England

February 7, 2005

Lancaster

- Lancaster also has a castle
- This castle is used as a prison

February 7, 20

Corpus Linguistics

- For the last 25 years, professors at Lancaster University have been conducting research in the area of computer corpus linguistics

February 7, 2005

Computer Corpus Linguistics

- Computer Corpus Linguistics is the sub-discipline of Computational Linguistics that utilises large quantities of texts to heighten the understanding of linguistic phenomena.
- A corpus is a collection of texts that has been assembled for linguistic analysis

February 7, 2005

Annotated Corpora

- Corpora are not much use to linguists in their raw form
- Annotated corpora are richer and more useful

February 7, 2005

Uses of Corpora

- Speech Synthesis
- Speech Recognition
- Lexicography
- Machine-aided Translation
- Information Retrieval
- EFL

Types of Annotation

- Grammatical Annotation
- Prosodic Annotation
- Syntactic Annotation
- Semantic Annotation

The_AT0 cat_NN1 sat_VVD on_PRP the_AT0 chair_NN1

[[S[NP *Pierre Vinken*],[ADJP[NP *61 years* old,]] *will*[VP *join*[NP *the board*][PP *as*][NP *a nonexecutive director*][NP *Nov 29*.]]]

Joanna	231112
stubbed	21072-31246
out	21072-31246
her	0
cigarette	2111014
with	0
unnecessary	317
fierceness	227052

231112	Personal Name
21072	Physical Activity
31246	Appearance
0	Low Content Word
2111014	Luxury Items
317	Causality / Chance
227052	Anger

Part-of-Speech Tagging

- POS tagging is the process of automatically assigning POS tags to words in running text
- An example:
  - Where are you from?
  - Where_AVQ are_VBB you_PNP from_PRP ?_PUN
- POS taggers have been developed for English and other Indo-European with varying degrees of success

Techniques Used in POS Tagging

- Statistical Techniques
- Rule-Based Techniques
- Machine Learning
- Neural Networks

Arabic Corpus Linguistics

- Arabic Corpora
  - No publicly available Arabic corpora
  - No Arabic POS tagger
- Arabic POS Taggers
  - POS taggers are being developed at New Mexico State University and Alexandria University

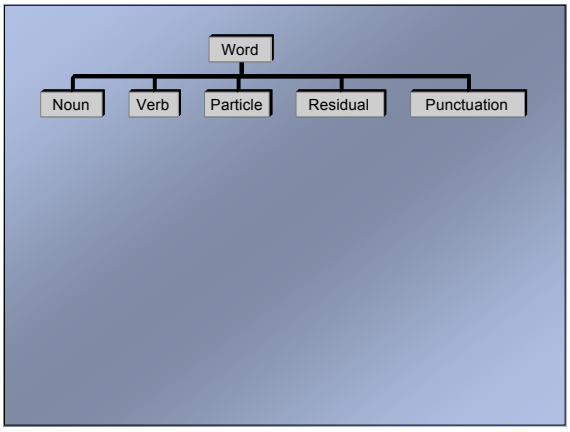
Developing the Arabic POS Tagger

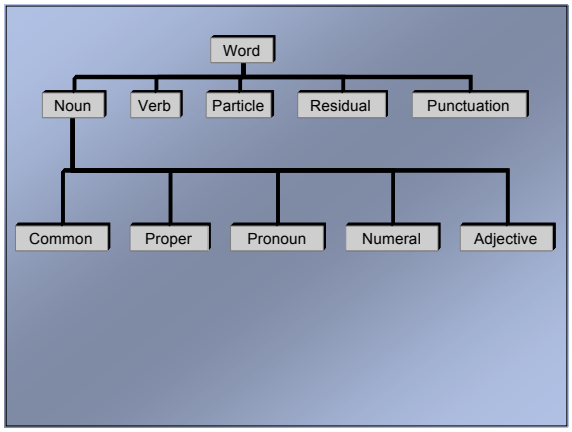
- Developing a manual tagger
  - Used to create a training corpus
- Defining an Arabic tagset
- Developing the automatic Arabic POS Tagger

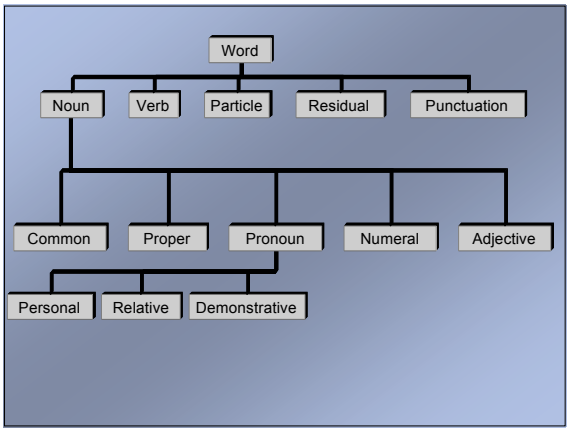
The Arabic Tagset

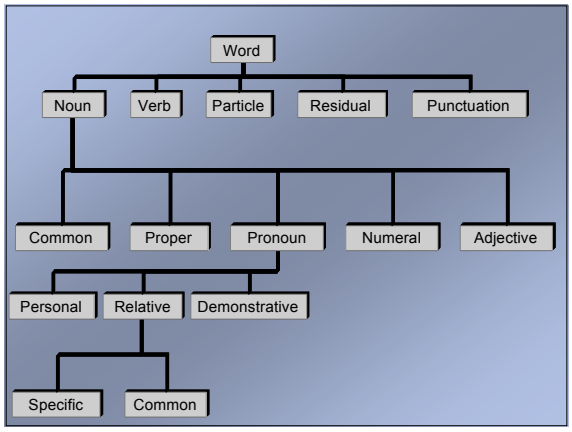
- Follows the tagging system that has been used for fourteen centuries
- Noun, verb and particle
- All subclasses of these three tags inherit properties from the parent class
- Many subclasses are described in Arabic such as nouns of instrument and nouns of place and time

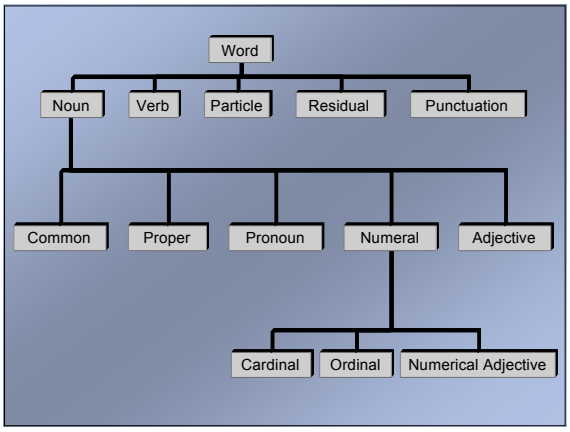
February 7, 2005

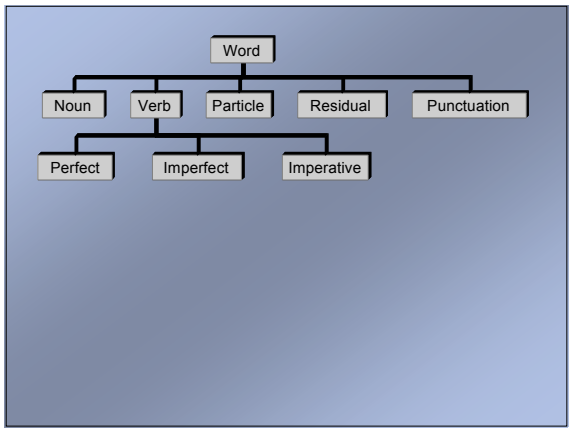


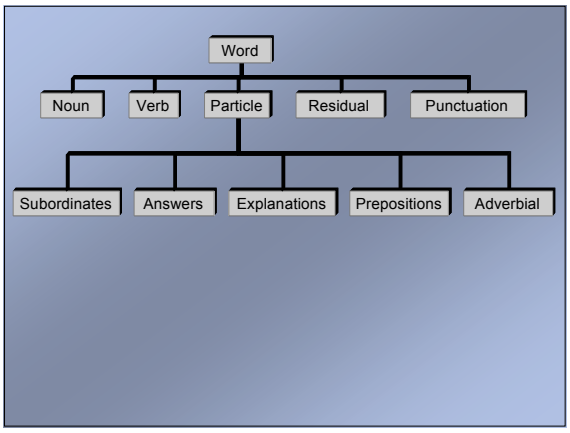


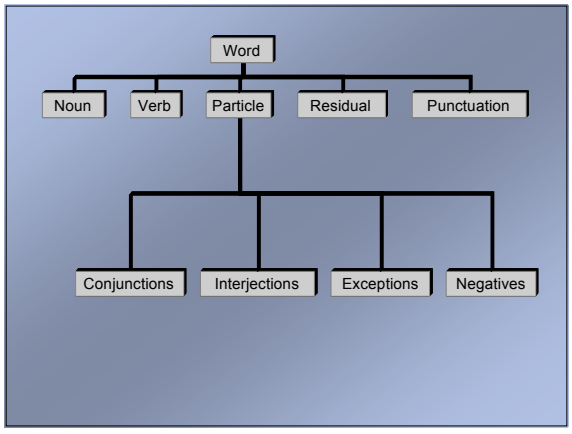


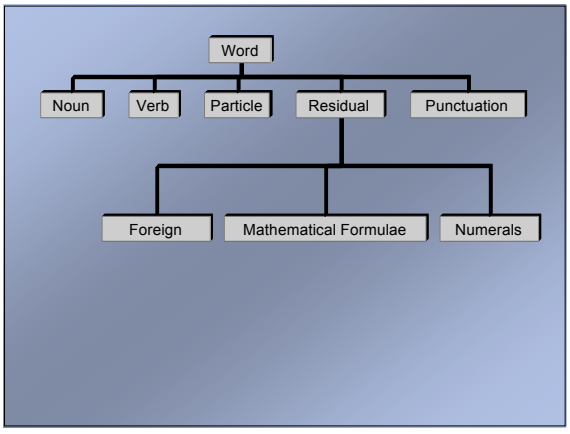


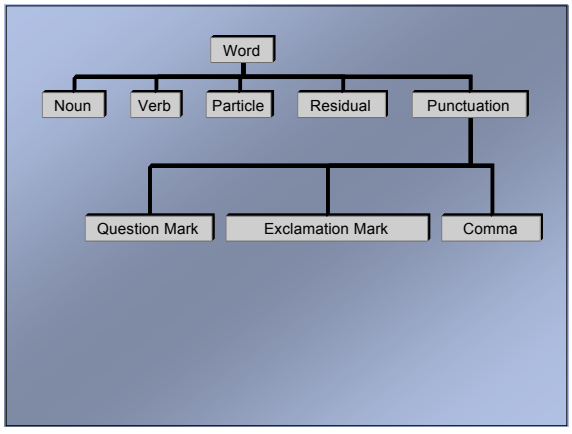


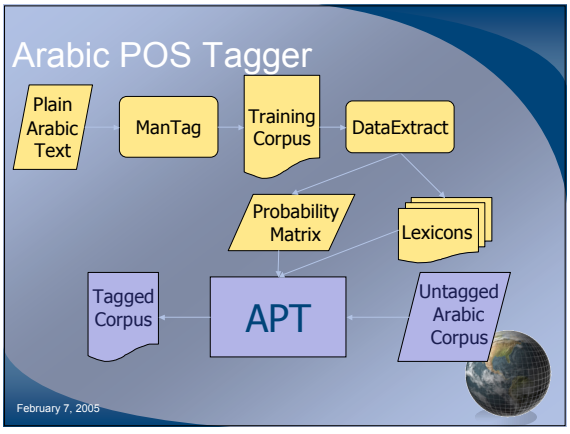

Word
Noun
Verb
Particle
Residual
Punctuation
Perfect
Imperfect
Imperative


Word
Noun
Verb
Particle
Residual
Punctuation
Subordinates
Answers
Explanations
Prepositions
Adverbial


Word
Noun
Verb
Particle
Residual
Punctuation
Conjunctions
Interjections
Exceptions
Negatives


Word
Noun
Verb
Particle
Residual
Punctuation
Foreign
Mathematical Formulae
Numerals


Word
Noun
Verb
Particle
Residual
Punctuation
Question Mark
Exclamation Mark
Comma


Arabic POS Tagger
Plain Arabic Text
ManTag
Training Corpus
DataExtract
Probability Matrix
Lexicons
Tagged Corpus
APT
Untagged Arabic Corpus
February 7, 2005

DataExtract Process

- Takes in a tagged corpus and extracts various lexicons and the probability matrix
  - Lexicon that includes all clitics.
    - (Sprout, 1992) defines a clitic as "a *syntactically* separate word that functions *phonologically* as an affix"
  - Lexicon that removes all clitics before adding the word

DataExtract Process

- Produces a probability matrix for various levels of the tagset
  - Lexical probability: probability of a word having a certain tag
  - Contextual probability: probability of a tag following another tag

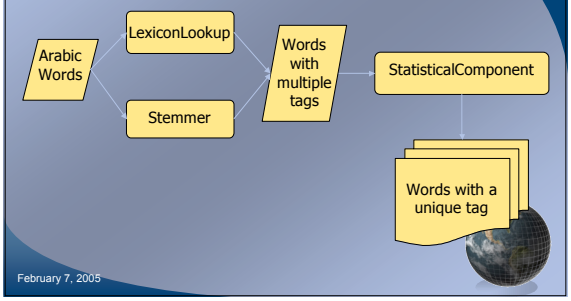
DataExtract Process

	N	V	P	No.	Pu.
N	0.711	0.065	0.143	0.010	0.071
V	0.926	0.037	0.0	0.008	0.029
P	0.689	0.199	0.085	0.016	0.011
No.	0.509	0.06	0.098	0.009	0.324
Pu.	0.492	0.159	0.152	0.046	0.151

Arabic Corpora

- 59,040 words of the Saudi ``al-Jazirah'' newspaper, dated 03/03/1999
- 3,104 words of the Egyptian ``al-Ahram'' newspaper, date 25/01/2000
- 5,811 words of the Qatari ``al-Bayan'' newspaper, date 25/01/2000
- 17,204 words of al-Mishkat, an Egyptian published paper in social science, April 1999

APT: Arabic Part-of-speech Tagger

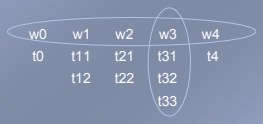


Result of Lexicon Lookup and Stemmer

- The Arabic word *fhm*
  - Would be assigned the tag of **conjunction and personal pronoun** by the morphological analyser to mean "so, they"
  - Would be found in the lexicon as the **noun** to mean "understand"
  - Would be found in the lexicon as the **conjunction and verb** meaning "so, he prepared"

Statistical Component

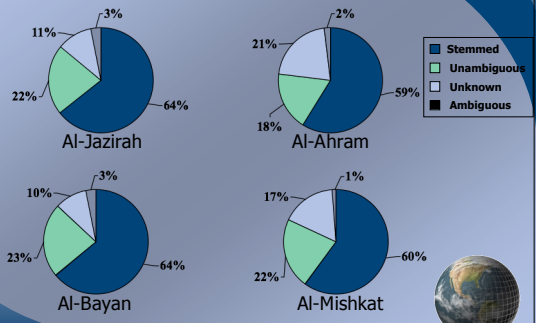
- Statistical Component.
  - Hidden Markov Models.



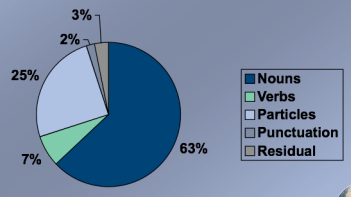
  - The probability of the words w0-w4 been tagged with t0, t11, t21, t31 and t4 is:

P(w0 | t0) * P(t11 | t0....) * P(w1 | t11) * P(t21 | t11....) * P(w2 | t21) * P(t31 | t21....) * P(w3 | t31) * P(t4 | t31....) * P(w4 | t4)

Initial Results



Other Findings



Distribution of Tags

Summary

- For my M.Sc. I developed an Arabic language stemmer
- For my Ph.D. I developed an Arabic part-of-speech tagger
  - The tagger used the stemmer

Advice

- Stay focused
- Do not procrastinate
- Make use of all the resources around you
- Write down everything
- More importantly, organise everything that you write down