Hybrid Method for Tagging Arabic Text

Written By: Yamina Tlili-Guiassa
University Badji Mokhtar Annaba, Algeria

Presented By: Ahmed Bukhamsin

Outline

- Introduction
- Overview of POS Tagging Techniques
- Hybrid Method For Tagging
- Rules-Based Tagging
- Memory-Based Learning
- Evaluation
- Results
- Conclusion

Introduction

- several important approaches to tagging
  - Hidden Markov Models
  - Finite StateTransducers

- Drawbacks of thease approches:
  - They are inflexible
  - Based on small amount of information

Introduction

- Approaches based on the position of the word in the sentence are not appropriate for tagging Arabic words.
  - Arabic has a weak positional constraint
  - Ambiguity in Arabic is enormous at every level
  - The absence of the short vowels increase the ambiguity

Overview of POS Tagging Techniques

- There are many methods which can be classified in three groups:
  - Linguistic approach
    - Based on set of rules written by linguists
  - Statistical approach
    - requires much less human effort
  - Machine learning based approach
    - Acquire a language model from a training corpus

Hybrid Method For Tagging

- Combining more than one method so it get the advantages of each one of them
  - Rules-based tagging
  - Machine learning based tagging

Rules-Based Tagging

- Affix signs
  - Proper to nouns
  - Proper to verbs
  - Proper to nouns and verbs
- The pattern signs
- Grammatical rules signs
- Other signs
  - Number
  - Gender
  - Preposition
  - Conjunction

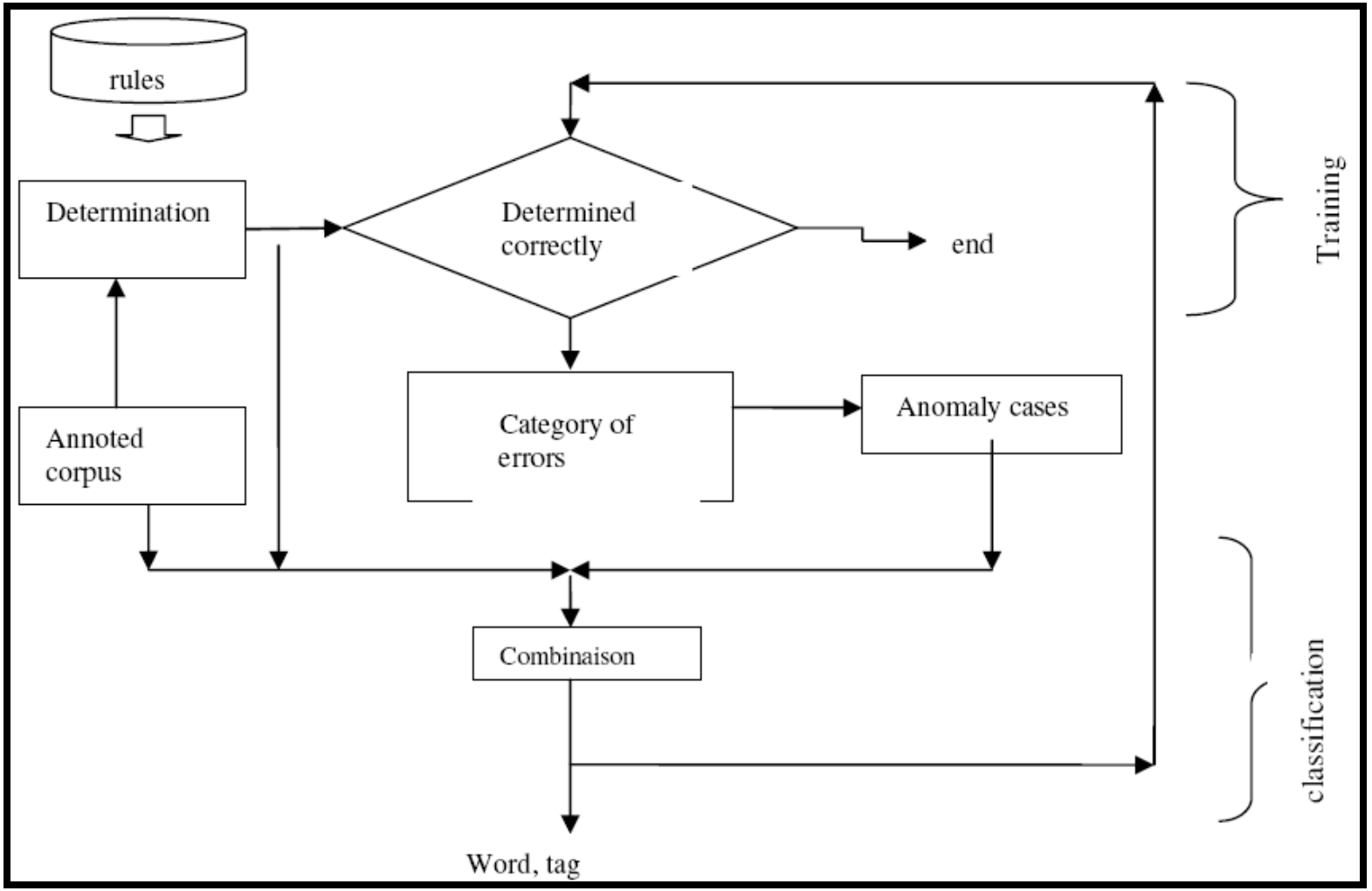
Memory-Based Learning

- Simple learning methods in where examples are massively retained in memory.
- The similarity between memory examples and new examples is used to predict the outcome of a new example.

Memory-Based Learning System

- Contains two components:
  - A learning component which is memory storage
  - A performance component that does similarity-based classification

Memory-Based Learning System



Evaluation

- Examples when only rules are applied:
- Example 1:
  - جَمِيْلٌ is a word with same consonant string and same vowels but has different tags: application of rule only produce the same tag for both cases.
  - جَمِيْلٌ يَشْرَبُ here جَمِيْلٌ must take the tag: NCSgMNI
  - الجَوُ جَمِيْلٌ adjective must take the tag: NACSgMNI

Evaluation

- Example 2:
  - دَخَلَتْ بِنْتٌ here بِنْتٌ is a noun but it take the tag: VPSg1
- Example 3:
  - شَتْان، هَيهَات …etc. show that a very high number of adjective can not be handled correctly and can be tagged as verbs.
- Example 3:
  - مدارس، أقلام، قصور and other broken plurals are classified as singular

Results

- Use of memory-based learning allows for easy integration of different information sources and can handle exceptions efficiently and has a number of advantages over statistical POS tagger.
  - Makes the tagging process more robust
  - Development time and processing is faster
  - Involves the disambiguation of word on basis of both sources

Results

- All experiments are performed on text extracted from educational books and some Qur'anic text. The tag set used is derived from APT.
- Rule based method gave 85% of correct result
- The Hyper method gave 98.2% of correct result

Results

- The figure shows some experimental results

Table 3: Results using rules-based and hybrid method

Test corpus	Rules only (%)	Rules only with correct pos tag (%)	Hybrid pos tag complete subtags (%)	Hybrid with correct complete subtag (%)
Originaltest	84.45	83.98	96.53	94.32
Test with pre-annoted names	88.06	86.48	98.01	97.00

Conclusion

- This proposed approach allows a new method for tagging Arabic by a combination of based-rules and a memory-based learning.
- This approach is based on linguistic rules and the tag is verified by memory-based learning.

Conclusion

- Rule-based system is quite easy to extend, maintain and modify.
- Such method combined with memory-based learning involved filling the gaps in the lexicon and modifying the POS tag set in order to meet the requirements of NLP tasks.
- The proposed approach can also be applied to other NLP processing such as chunking.

Thank you for listening

- Any Question ?