Applications of NLP

Shadi Saleh
Shamra's CEO

NCD - Latakia, DEC 2015

About me

- PhD student (Cross lingual information retrieval).
- CEO&Founder, Shamra Search engine.
- IOI 2011 (Thailand) 2012 Italy.
- ACM ICPC 2008-2013, World Finals Russian St. Petersburg and Ekaterinburg

What is NLP?

- Natural Language processing.
- Analysing and Parsing for our “Natural” languages
- Let the machine understand our spoken language??

Part of Speech Tagging

- Part of Speech Tagging
  - Zein gave her a flower
- Assign (tag) to each word in a given sentence
- It solves conflict cases, e.g:
  - I cut the telephone line
  - I walk the line
  - I will **book(?)** the ticket to Syria, then I will read the **book(?)**

POS example

- Parsing the following sentence using Stanford Arabic Parser:

Input: هذا الرجل هو سعيد.

Output: هذا/DT الرجل/DTNN هو/PRP سعيد/NNP ./PUNC

Tag-set: DT: Determiner, PRP: Personal pronoun, NNP: Proper noun, singular, PUNC:Punctuation

Famous Tag-set

- PENN POS
- Stanford POS
- Prague POS
- Moscow
- …??
- https://ufal.mff.cuni.cz/hamledt

HArmonized Multi-LanguagE Dependency Treebank

- HamleDT developed at Charles university in Prague
- Contains 42 treebanks integrated in HamleDT at this moment.

Word2Vec

- Word2vec is a two-layer neural net that processes text.
- Its input is a text corpus and its output is a set of vectors
- It detects similarities between words mathematically
- Given enough data, usage and contexts, Word2vec can make highly accurate guesses about a word's meaning based on past appearances.

Word2Vec

- A well trained Word2Vec can answer such question:
  - Exclude the word that does not belong to list:

1. math shopping reading science
2. eight six seven five three owe nine
3. More: https://github.com/dhammack/Word2VecExample

Now you guess!

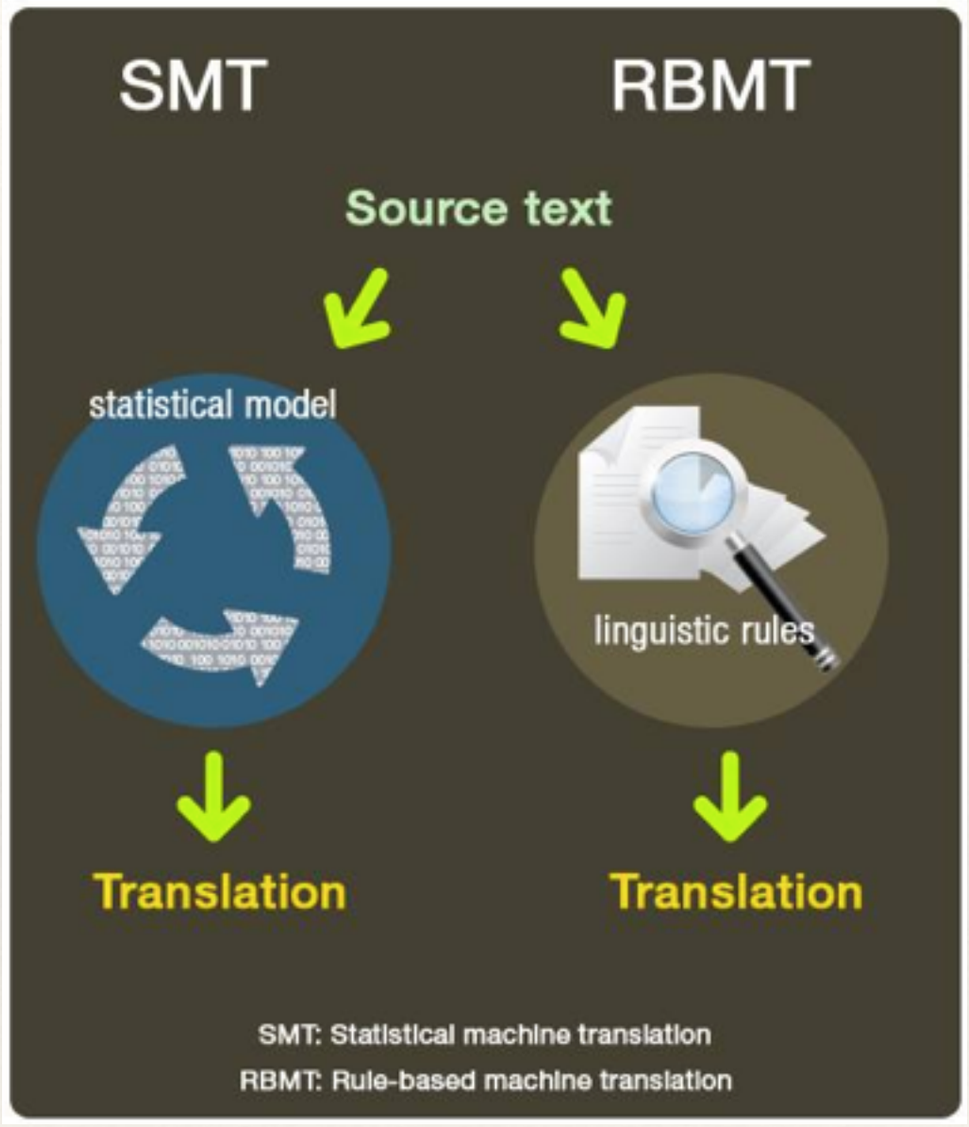
1. Obama + Russia - USA = ?
2. King - man + woman = ?
3. Library - Books = ?
4. President - Power = ?

Now you guess!

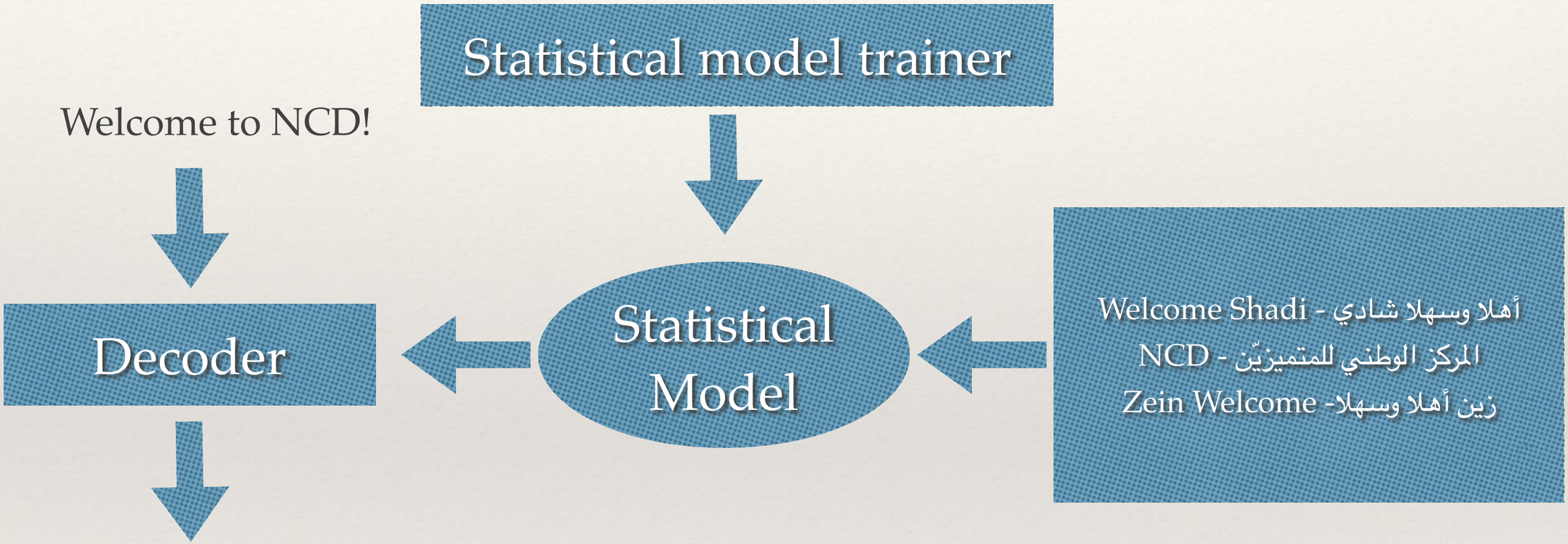
1. Obama + Russia - USA = Putin
2. King - man + woman = Queen
3. Library - Books = Hall
4. President - Power = Prime minister

Machine Translation Systems

- Mainly Two methods: SMT and RBMT
- SMT is used now (e.g Google Translate)
- Can be word-based or Phrase-based



SMT architecture



Arabic Text Tashkeel

- Apply Statistical approaches to achieve such task

هَلْ تَعْلَم أَنَّ هُنَاكَ أَكْثَر مِنْ ٢٠٠ مِلْيُون شَخْصٍ يَدْرِسُونَ اللُّغَةَ العَرَبِيَّةَ وَهِيَ لَيْسَتْ لُغَتهُم الأُمِّ؟ أَحَدُ المُشْكِلَاتِ الَّتِي تُوَاجِهُ هَؤُلَاءِ الدَارِسِينَ عِنْدَ مُحَاوَلَةِ القِرَاءَةِ باللُّغَةِ العَرَبِيَّةِ هِيَ عَدَمُ وُجُودِ عَلَامَاتِ التَشْكِيلِ وَهُوَ مَا يَجْعَلُ فَهْمَ الكَلِمَةِ وَنُطْقَهَا نُطْقاً صَحِيحاً أَكْثَرُ صُعُوبَةً.

Semantic similarity

- How much two sentences are similar?
- Here are the steps for computing semantic similarity between two sentences:

1. Each sentence is partitioned into a list of tokens.
2. Part-of-speech (or tagging).
3. Stemming words.
4. Find the most appropriate sense for every word in a sentence (Word Sense Disambiguation)
5. Compute the similarity of the sentences based on the similarity of the pairs of words.

Social Sentiment Analysis

- Sentiment analysis in social media allows us to get an idea about public opinion regarding certain topics.
- Can be binary (1,0) or multipolar:

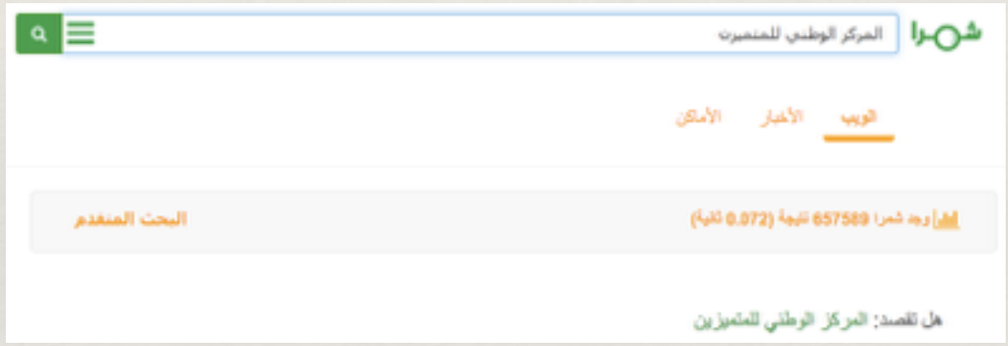
- Do you know that some researchers predicted the terrorism attack in Tunisia one week before using SSA?

More Applications …

- Named Entities recogniser (NER)
- Information Retrieval (IR)
- Text simplification and summarisation
- Paraphrasing
- Question/Answering system
- Dialogue system

NLP in Shamra!

- Autocomplete
- Autocorrection
- Knowledge search
- Advanced methods used in Search optimising and query normalising



Thank for your kindly attention!

Questions?