NLP and IR

Course Projects

Aliaksei Severyn

University of Trento, Italy

March 26, 2013

Projects

- Question Classification
- Question Focus detection
- Wikification
- Semantic Textual Similarity
  - Semantic features from WikiPedia
  - Advanced syntactic/shallow semantic trees
  - Exploring SRL for semantic similarity
  - Sentence type identification

Question Classification

Task: classify a question into one of the pre-defined categories

DESC:manner How did serfdom develop in and then leave Russia ?

ENTY:cremat What films featured the character Popeye Doyle ?

DESC:manner How can I find a list of celebrities ' real names ?

ENTY:animal What fowl grabs the spotlight after the Chinese Year of the Monkey ?

ABBR:exp What is the full form of .com ?

HUM:ind What contemptible scoundrel stole the cork from my lunch ?

HUM:gr What team did baseball 's St. Louis Browns become ?

HUM:title What is the oldest profession ?

Question Classification: demo

```
sovarm@dhcp089:~/data/biomed/qapipeline$ sh demo_question_classifier.sh
*** Question classifier demo ***
--
Please type a question:
> Who was the inventor of silly putty?
The predicted question class is: HUM
--
Please type a question:
> Where is Trento located?
The predicted question class is: LOC
--
Please type a question:
>
```

Question Focus Detection

Task: detect the key word in the question that directly points to what is being asked.

Name 11 famous #martyrs .

What 's the Olympic #motto

What is the #origin of the name Scarlett '

What 's the second-most-used #vowel in English

Who was the #inventor of silly putty

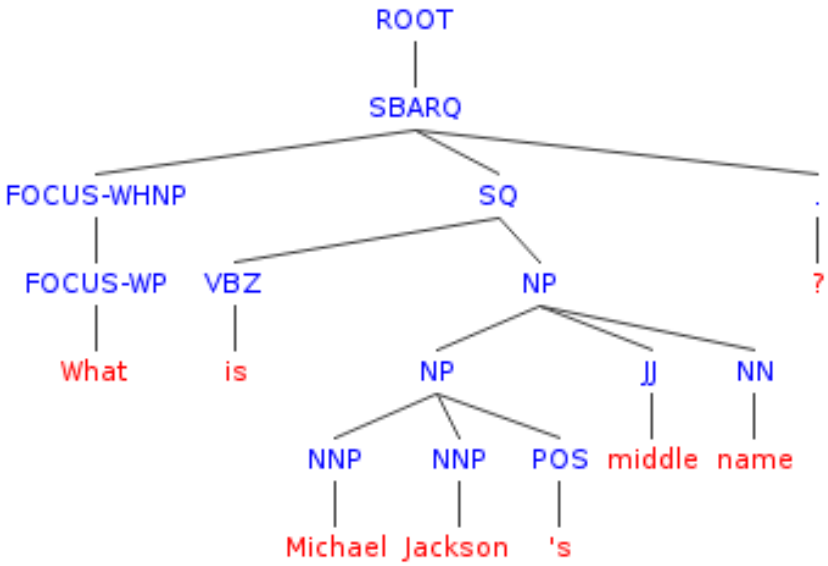
What is the highest #waterfall in the United States

Name a golf #course in Myrtle Beach .

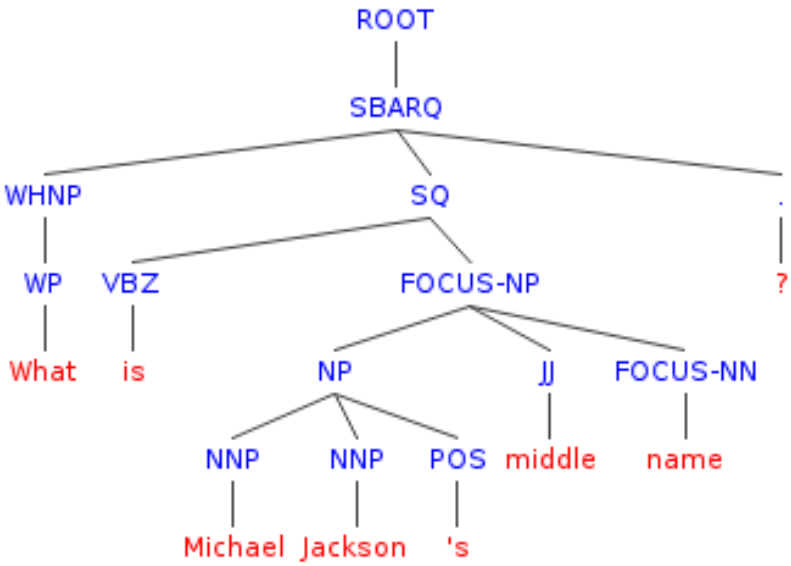
Focus detection: Demo

```
sovarm@dhcp089:~/data/biomed/qapipeline$ sh demo_focus_classifier.sh
*** Question focus demo ***
--
Please type a question:
> Who was the inventor of silly putty?
Who was the inventor of silly putty? | Focus: inventor
--
Please type a question:
> Where is Trento located?
Where is Trento located? | Focus: Trento
--
Please type a question:
>
```

Focus detection: Tree Representation



Incorrect candidate



correct candidate

Wikification

Task: identify important entities and concepts in text, disambiguates them and link them to the related Wikipedia concepts.

Houston, **Monday**, July 21 -- **Men** have landed and walked on the moon. Two Americans, astronauts of Apollo 11, steered their fragile four-legged lunar module safely and smoothly to the historic landing yesterday at 4:17:40 P.M., Eastern daylight time. Neil A. Armstrong, the 38-year-old **civilian commander**, radioed to earth and the mission control room here: "Houston, Tranquility Base here; the Eagle has landed."

The first **men** to reach the moon -- Mr. Armstrong and his co-pilot, Col. Edwin E. Aldrin, Jr. of the Air Force -- brought their **ship** to rest on a level, **rock-strewn plain** near the southwestern shore of the arid Sea of Tranquility. About six and a half **hours** later, Mr. Armstrong opened the landing craft's hatch, stepped slowly down the ladder and declared as he planted the first human footprint on the lunar crust: "That's one small step for man, one **giant** leap for mankind."

Semantic features from Wikipedia

Task: Use Wikipedia to define semantic similarity features for a pair of text snippets, i.e. sentences.

Ideas:

- Similarity based on top concepts for a given sentence returned from the Search Engine
- Exploiting Wikipedia categories to establish the matching

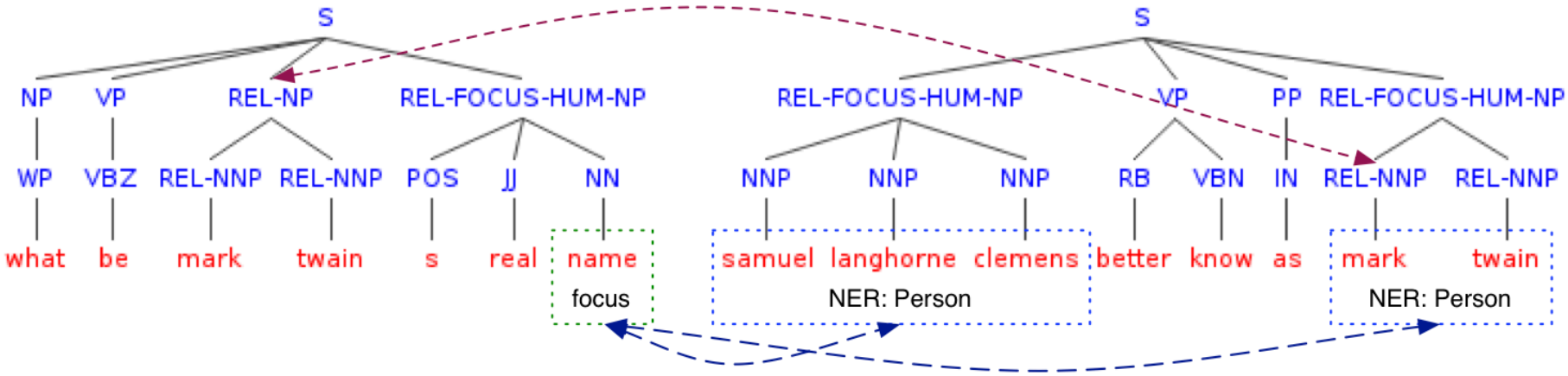
ESA demo

```
rm@dhcp089:/mnt/sdb/sovarm/descartes-0.2$ sh esa-demo.sh 10 questions.txt                    [146/764
4:59.859 [main] INFO  e.i.c.c.d.retrieval.simple.Searcher - Opened index located at data/wiki-index
4:59.862 [main] INFO  e.i.c.c.d.retrieval.SearcherFactory - Created searcher with the following configurat

    Index directory: data/wiki-index,
    Use text and titles: false,
    Use ngrams: true,
    Stopwords: [to, com, for, www, about, by, from, where, I, who, was, de, of, are, when, on, be, with, i
how, or, a, at, as, the, in, und, that, what, an, will, la, en, this]
 In what country did the game of croquet originate?
    Croquet
    Jack Osborn
    NCAA Men s Division I Basketball Championship
    Extreme croquet
    Bowl Championship Series
    History of American football
    All England Lawn Tennis and Croquet Club
    Eglinton Country Park
    Zambia women s national football team
    Ben Rothman
 Who is Tom Cruise married to?
    Tom Cruise
    Tom Cruise: An Unauthorized Biography
    Valkyrie (film)
    Nicole Kidman
    Minority Report (film)
    Knight and Day
    Cruise ship pollution in the United States
    Being Tom Cruise
```

Advanced syntactic/semantic trees

Task: Augment constituency/dependency parse trees with some kind of semantic annotations, e.g. WordNet class/hypernyms/ etc., NERs, SuperSense tags.



Exploring SRL for semantic similarity

Task: Use SRL to construct shallow semantic structures for a given sentence + tree kernels to generate the feature spaces.

In	IN	O	-	O	B-AM-LOC
what	WP	O	-	B-R-A0	I-AM-LOC
country	NN	O	-	E-R-A0	E-AM-LOC
did	VBD	O	did	S-V	O
the	DT	O	-	B-A1	B-A1
game	NN	O	-	I-A1	I-A1
of	IN	O	-	I-A1	I-A1
croquet	NN	O	-	I-A1	E-A1
originate	VBP	O	originate	E-A1	S-V
?	.	O	-	O	O
Who	WP	O	-	S-R-A1	
is	VBZ	O	-	O	
Tom	NNP	B-PER	-	B-A1	
Cruise	NNP	E-PER	-	E-A1	
married	VBD	O	married	S-V	
to	TO	O	-	O	
?	.	O	-	O	

Sentence type classifier

Task: Perform clustering/classification to learn the sentence types. Explore different sets of lexical/syntactic/semantic features to learn an accurate classifier. The obtained features can then be pluggged into the final classifier. Use additional corpora to learn useful priors.

Other ideas

- Pay attention during the course
- To come up with an idea you would like to work on for the course project
- Consult with Alessandro or me

Resources

- More detailed description here:

  http://goo.gl/sC8ah

- Previous year projects:

  http://goo.gl/gPJ9D