Arabic Community Question Answering

Yonatan Belinkov, Mitra Mohtarami, James Glass, MIT-CSAIL, Cambridge, MA 02139, USA
Shafiq Joty, Alessandro Moschitti, Preslav Nakov, Fahad A. Al Obaidli, QCRI, HKBU, Qatar

Background

- Community forums where user post questions and answers are becoming increasingly popular
- Arabic community question answering (cQA) received little attention in the past
- The SemEval-2016 Task 3 offered a cQA Arabic dataset in the medical domain, where:
  - Given a question and 30 related question-answer pairs, rank the pairs with respect to the original question
  - QA pairs annotated as *Direct*, *Relevant*, or *Irrelevant*
- Challenges: long texts, user-generated content, medical terminology, mixed standard and colloquial language
- Example:
  - Q': ما هي اعراض الاصابة بمتلازمة القولون المتهيج العصبي
  - Q: ما هي أعراض القولون العصبي
  - A: تتباين علامات وأعراض الاصابة بمتلازمة القولون العصبي بين الأشخاص [. . .] كالتالي : – ألم أو تشنج البطن . – الشعور بالانتفاخ . – امتلاء البطن بالغازات . – الاسهال أو الامساك [. . .]

Feature Representation

- Given a question *q'* and a related question-answer pair *q-a*, compute features between the pairs *q'-q* and *q'-a*
- **Text-based features**
  - Various text-similarity metrics such as Longest Common Substring, Longest Common Subsequence, Greedy String Tiling, etc. (Belinkov et al. 2015)
- **Vector-based features**
  - Vector representations of closest pairs of words or sentences in *q'-q* and *q'-a*
  - Word vectors computed from Arabic Gigaword and medical domain raw data using *Word2Vec*
  - Sentence representation is average of word vectors
- **Machine translation evaluation features**
  - BlEU, TER, Meteor

Preprocessing with Keyword Extraction

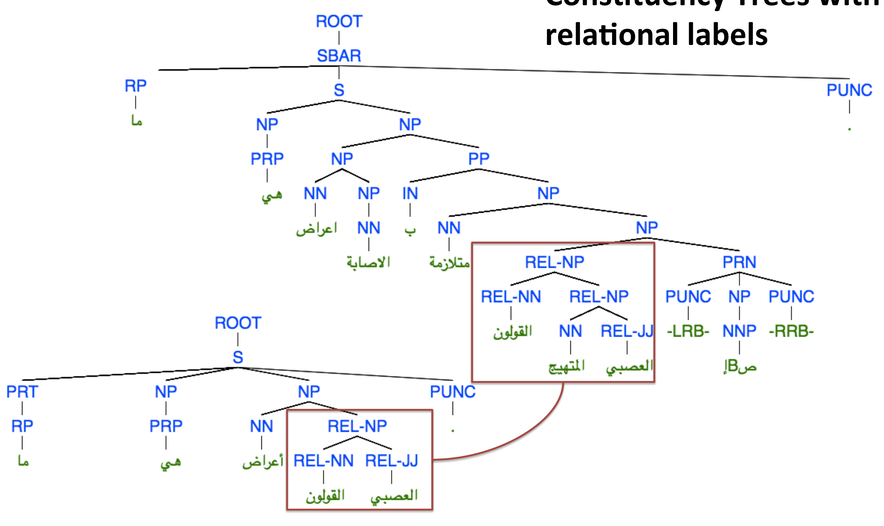
- **Text preprocessing** in order to deal with several challenges:
  - Long texts: average question/answer length is 50/120 words
  - Rich morphology: multiple surface forms per lemma
  - Latin terminology in the medical domain
- **Keyword extraction** with TextRank
  - Treat every thread as a document
  - Form a graph where nodes are word types and edges represent co-occurrence in *N*-sized window
  - Compute importance weight iteratively and keep top *P%* of words
- **Lemmatization**: we apply MADA for finding lemmas and part-of-speech tags
- **Stop-word removal**: we keep only content words, Latin words, and words with no morphological analysis.
- **Sentence Splitting:** We use the splitter from Stanford's core NLP toolkit.
- **Syntactic parsing:** We use the Berkeley parser to get the syntactic tree for each each sentence.

Tree Kernels

Syntactic Tree Kernels

$$K((t_1, t_2), (u_1, u_2)) = TK(t_1, u_1) + TK(t_2, u_2)$$

Constituency Trees with relational labels



Experiments

- **Preprocessing settings**
  - i. No preprocessing
  - ii. Only keeping content lemmas
  - iii. Only content lemmas and keyword extraction with TextRank params *N=3*, *P=5*
  - iv. Same, with TextRank params *N=4*, *P=1*
- **Tree kernels settings**
  - a. ConvKN-contrastive1: only basic features
  - b. ConvKN-contrastive2: MT features
  - **c. ConvKN-primary**: basic features + tree kernels

	MAP	AvgRec	MRR	P	R	F1	ACC
Rand	29.79	31.00	33.71	19.53	20.66	20.08	68.35
a	38.33	42.09	43.75	20.38	**96.95**	33.68	26.58
b	39.98	43.68	46.41	26.26	68.39	37.95	57.00
c	**45.50**	**50.13**	**52.55**	**28.55**	64.53	**39.58**	**62.10**
i	44.94	49.72	51.58	**62.96**	2.40	4.62	**80.95**
iii	42.95	47.61	49.55	27.20	**74.40**	39.84	56.76
i-iv	**45.83**	**51.01**	**53.66**	34.45	52.33	**41.55**	71.67

Future Work

- Combine keyword extraction with tree kernels
- How do deal with grammatical structure after keyword extraction?
- Automatically detecting the most important sentences to be matched with the tree kernels

References

- Belinkov et al. 2015. VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems.
- Mohtarami et al. 2016. SLS at SemEval-2016 Task 3: Neural-based Approaches for Ranking in Community Question Answering.
- ConvKN at SemEval-2016 Task 3: Answer and Question Selection for Question Answering on Arabic and English Fora