



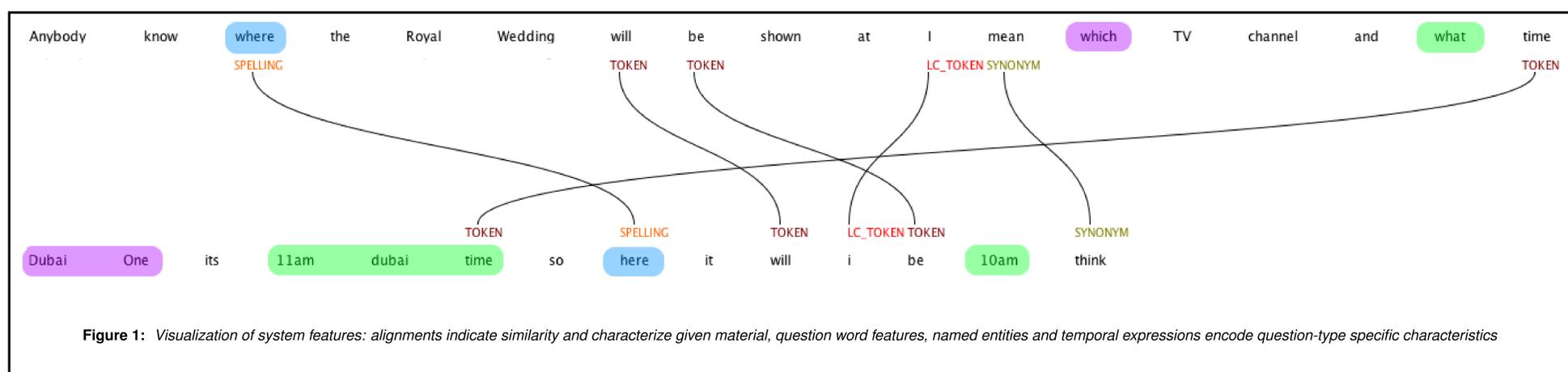
Introduction

- **Setting:** QA scenario where information retrieval is completed and answer candidates need to be assessed
- **Goal:** select answers that potentially help answering a given question.
- **Our background:** classification of learner answers to questions by comparing them to reference answers
- **Approach to Answer Selection:**
 1. Use our Short Answer Assessment system to compute similarity between question and answer.
 2. Extract additional features from answer part not aligned by step 1.

System Details

Feature	Description
<i>CoMiC Features</i>	
1. Keyword Overlap	% dependency heads aligned (relative to question)
2./3. Token Overlap	% aligned question/candidate tokens
4./5. Chunk Overlap	% aligned question/candidate syntactic chunks
6./7. Triple Overlap	% aligned question/candidate dependency triples
8. Token Match	% token alignments that were token-identical
9. Similarity Match	% token alignments resolved via PMI-IR
10. Type Match	% token alignments resolved via WordNet
11. Lemma Match	% token alignments that were lemma-resolved
12. Synonym Match	% token alignments sharing same WordNet synset
13. Variety of Match (0-5)	Number of kinds of token-level alignments (feat. 8–12)
<i>New Features</i>	
14–18. POS Weighting	Relative frequency of tokens from POS classes <i>noun, verb, adj/v, rest</i> in unaligned material
19–30. Question Words	Presence of <i>who, how, why, when, where, which, whom, whose, what</i> and modals or auxiliaries in question
31–34. Named Entities	Presence of PERSON, ORGANIZATION, LOCATION named entity in new answer material as indicated by Standard NER (Finkel et al., 2005)
35. Num Chunks	Number of syntactic chunks in answer
36. Temporal Expressions	Presence of temporal expressions in answer as found by HeidelTime tagger (Strötgen and Gertz, 2013)

- CoMiC features measure general similarity between question and answers on different linguistic levels.
- More specific relations between question and answers modeled with question words and new answer features
- Noisy web content (emojis, hashtags, ...) excluded from analysis with ArkTweet parser (Gimpel et al., 2011)
- The features above are exemplified in Figure 1 (below), with corresponding elements in the same colors.



Data

Data	Questions	Answers	Good	Bad	Potential	Dialogue	non-English	Other
Dev.	300	1645	53.1%	16.3%	11.3%	18.9%	1.0%	0.0%
Train.	2600	16,541	48.7%	18.0%	10.0%	22.7%	0.4%	0.0%
Test	329	1976	50.5%	18.3%	8.4%	22.0%	0.7%	0.0%

- Questions and labeled responses from Qatar Living web forum (Màrquez et al., 2015)
- Each response with label indicating the potential for answering the question:
 - *Good, Bad, Potential, Dialogue, non-English, Other*
 - uneven distribution of label frequencies
- Informal English strongly deviating from standard
- Example from the training data:

Question: "TO all Filipinos here in DOHA. Would you like to have the ZAGU (black pearl shake) here in DOHA? Just "YES" or "NO" Please."
Answer (Good): "Oh my Yhes!!!! :)"
Answer (Bad): "Although I love Quickly better... :(""
Answer (Dialogue): "so where is zagu now?"

Results

System	Development Set		Test Set	
	Accuracy	Macro F1	Accuracy	Macro F1
Best system	–	–	73.76%	57.29%
CoMiC-primary	54.89%	28.41%	54.20%	30.63%
CoMiC-contrastive	53.37%	24.36%	50.56%	23.36%
Majority baseline	53.19%	23.15%	50.46%	22.36%

- Primary: CRF approach (MALLETT, McCallum 2002) with Markov order 1 to classify a *sequence* of answers
- Contrastive: memory-based learning (TiMBL, Daelemans et al. 2007) using cosine distance and $k = 5$
- CRF performs better, probably since it accounts for the conversation character by using an answer's context.

Conclusion

- We applied an alignment-based Short Answer Assessment system to the task of Answer Selection,
 - using question-answer alignment to determine whether answers are on-topic, and
 - characterizing new, unaligned content in answers on different linguistic levels.
- Modest performance, mainly due to insufficient alignments found in the often ungrammatical data.
- First step in developing fine-grained question-answer features for our main task, Short Answer Assessment.