Winograd Schema Challenge with Narrative Chains

Sung Su Lim

Computer Science and Engineering

The Hong Kong University of Science and Technology

Abstract

In this project we examine and analyze Winograd Schema Challenge, a set of challenging questions on resolving pronouns, also referred to as a coreference task. We tackle a small domain in this set of questions, where the two actors are involved in a single event, and the pronoun is involved in another. We attempt to resolve the pronoun through Narrative Chains. Our results show that despite some of its limitations, it may be an effective way to resolve pronouns given enough data. We further extend the work into examining future improvements by using word embeddings.

Introduction

The Winograd Schema Challenge is also often referred to as an alternative Turing Test. It is a set of questions specially designed so that it is extremely easy for humans with commonsense knowledge to resolve the pronouns, but difficult for machines without commonsense. One of the hardest challenge is in tackling this problem is to encode real-world knowledge, since such knowledge is shown to be crucial in resolving the pronouns.

The following is a pair of example from the Winograd Schema Challenge:

*The city councilmen refused the demonstrators a permit because **they** (feared/advocated) violence.*

A. The city councilmen

B. The demonstrators

The question asks to indicate what the pronoun 'they' refers to. The two sentences differ by a single word, yet the answer changes completely. It is important for the machine to capture such subtle difference, and the question is designed so that it cannot be resolved by other trivial knowledge (i.e. singular/plural forms or grammatical structure). Both answers make perfect sense grammatically. Thus the machine has to 'reason' with real-world knowledge that *'the city councilmen refused the demonstrators a permit'* **because** *'the city councilmen feared violence'* **or** *'the demonstrators advocated violence'*.

In this project we attempt to tackle a small domain of sentences, specifically sentences that contain (1) one main event in which the two answer choices are each involved as subject and object respectively, (2) the pronoun is involved in another event acting as the subject or the object. We extract a binary feature using **Narrative Chains**, which will be explained further in the later sections.

Narrative Chains

Narrative Event Chain is a set of events centered on a single common protagonist, in order to encode some knowledge of related events (actions). They are similar to scripts (Schank and Abelson, 1977), which are handwritten structured sequences of participants and events in order to capture the knowledge of a certain small domain. Chambers and Jurafsky propose 'Narrative Event Chains' as an unsupervised way of learning to cluster such related events together from a raw news text. We use narrative chains of length 12[1], which is the largest narrative chain dataset available.

The following is an example of the Narrative Chain(Chambers and Jurafsky, 2008):

[force-s, harass-s, bully-s, sue-o, coerce-s, assault-s]

A Narrative Chain is composed of a number of related events(actions) and the role of the protagonist. Here '-s' denotes a subject role, and '-o' denotes an object role. We see that someone who 'forces' or 'bullies' is also likely to 'be sued' or 'assault'. Using this extracted knowledge of events that are related to each other, we attempt to apply it to resolving pronouns.

[1] The narrative chain dataset can be found in https://www.usna.edu/Users/cs/nchamber/data/schemas/acl09/

Preprocessing the Sentence

We first preprocess the sentences to allow easier feature extraction. Using Stanford CoreNLP semantic role-labeler, we parse the sentences and extract semantic components and their relationships. We also simplify the answer candidates in the sentence by replacing the noun phrase with the main noun-head, i.e. ‘*the brown suitcase*’ would be replaced with ‘*suitcase*’. Furthermore, we found that there were questions with more than one sentence. As a remedy we concatenate the two sentences with the connective ‘and’.

Feature Extraction

We extract the NC (Narrative Chains) feature following the methods proposed in Rahman and Ng (2012):

1. Given a sentence, we first identify the main event E_m that involves the two answer candidates, and determine the role of each candidate. Each candidate’s role-event would be E_m-*s* or E_m-*o*.

2. We then determine the event that the pronoun participates in, and its role.
$$E_p\text{-}k \quad where \;\; k \in \{s, o\}$$

3. We pair the pronoun event-role with two possible main event-roles. *(E_p-k, E_m-s)* and *(E_p-k, E_m-o)*

4. We search for the occurrence of these two pairs in the Narrative Chain corpus.

5. If such pair is found, we identify the role of the main event, and match the answer according to the candidate’s event-role.

More specifically, given the example:

*Joe paid the detective after **he** received the final report on the case.*

A. Joe

B. the detective

The main event would be 'pay' and the corresponding roles of the candidates would be 'pay-s' and 'pay-o' respectively. The event-role that the pronoun 'he' participates in would then be 'receive-s'.

Our program outputs:

```
main_event_root: pay

A: paid-nsubj

B: paid-dobj

pron_event_list: ['receive']-nsubj

pron_event_type_list: ['VERB']

found:[u'pay-s', 'receive-s']
```

Note that our program generates the pronoun event in a list. It is because in some sentences like

‘Bob killed Vincent because he wanted to kill someone.’

the pronoun is linked with two possible events: ‘*want*’ and ‘*kill*’. In such cases we search for both pronoun events.

Experimental Results

We test the performance of our binary NC feature on three different datasets. We did implement Rank SVM to use our binary NC feature, but we believed it was trivial because we only utilize one feature with binary values. The SVM seemed to be predicting randomly when there weren’t any NC features extracted from the corpus. We match the answer with the binary output value of the NC feature.

Dataset

In addition to the 272 questions provided by the Winograd Schema Challenge, we further found two more datasets by Rahman and Ng (2012). These datasets were collected from students with the same constraints as in the Winograd Schema Challenge (WSC). Rahman and Ng (2012) provide a training set composed of 1322 questions and a test set of 564 questions[2], which we denote as HLTRI_train and HLTRI_test.

Results and Discussions

The following table shows the number of instances classified to each category. We found that a large majority of the questions did not satisfy the constraints set forth by us in the Introduction and results in Not Found.

[2] The training and testing set can be found in http://www.hlt.utdallas.edu/~vince/data/emnlp12/

	Correct	**Incorrect**	**Not Found**
WSC	3	1	272
HLTRI_train	25	25	1272
HLTRI_test	17	9	538
Total	**45**	**35**	**2082**

There are several causes of ‘Not Found’: 1) The Stanford Parser itself contains some errors, 2) the sentence structure does not meet our constraints and is not parsed correctly, and 3) the identified NC event-role pair is not found in our corpus.

The first issue is concerned with the preprocessing of the questions. Semantic Role Labeling and Parsing are fields of active research and there are some uncertainties in accurate role labeling. In the sentence *‘(the) sniper shot (the) terrorist because he was a bad guy’*, instead of identifying *‘shot’* as a verb, it classifies *‘sniper shot’* as a compound noun.

The second issue is often related with the complexity of the given question sentences. In some sentences the answer candidates are in the form of possessive nouns (i.e. Ann’s son), while in others the answer candidates are not directly involved in the same main event.

The third issue arises because of the limited scope of our corpus. Many of the events are not included in the corpus; the data that the corpus had been extracted from may not have had enough exposure to those events to extract such Narrative Chains. Another limitation with the corpus was that it only handles ‘events (verbs)’. *‘Mary cleaned Susan's room and she was thankful.’* Here, the pronoun is related to the event ‘*was thankful*’. If the adjective ‘*thankful*’ could be incorporated to the corpus, it would also improve the accuracy.

For ‘incorrect instances’, we found that there were some sentences with the same verb for ‘main event’ and ‘pronoun event’: *‘The chimpanzee could not use Linux because it uses different commands than Windows.’* In other cases, event-roles were extracted correctly but was classified incorrectly. We suspect it is because we had not incorporated any knowledge from the connectives (because, although) or negatives (not) that could have inverted the meaning of the events.

Needless to say, among the small domain of sentences that our NC features were found in the corpus, it achieved 56% in accuracy. With further modifications to our NC feature extraction and addition of more features, as discussed in Rahman and Ng (2012), we believe it will achieve a much better result.

Future Work

From our insights from the NC feature, we attempted to extend our work further by analyzing the similarity of words. From our research, there were two main approaches.

One approach is using word embeddings, where a word is represented as a point in a 300-dimensional feature space. This word embeddings (T. Mikolov et al. 2013) are very commonly used in NLP to determine the similarity of word contexts. In the famous example, from a well-trained word embedding feature space, the closest word that matches the operation ‘king’ - ‘man’ + ‘woman’ is ‘queen.’

Preliminary investigation reveals that the similarity of the words ‘*refuse*’, ‘*fear*’, ‘*advocate*’, and ‘*violence*’, Sim(refuse, fear) = 0.34148, Sim(refuse, advocate) = 0.08198, and Sim(refuse, violence) = 0.10318. From the first example introduced in the introduction, this word similarity may be useful in that ‘*refuse*’ and ‘*fear*’ are more similar than ‘*refuse*’ and ‘*advocate*’. And since the events are in active voice, it could be correctly matched to the first answer candidate ‘*the councilmen*.’ There also needs to be further consideration with regards to other factors such as connectives, causal relations, and complements that could change the relationship between the two words.

Another approach is using the similarity of ‘word senses.’ A single word can have several meanings in different contexts, and word sense is one of the meanings of the word. Hence, it can capture the true context of the word, given that this word sense is correctly identified. In WordNet, the lexical database of English, the word senses are grouped into sets of synonyms called SynSets.

Word embeddings and word senses use word context knowledge similarly as with the Narrative Chains corpus. Due to time constraint we could not explore further, but incorporating such state-of-the-art approaches pronoun resolution may achieve better results.

[2] The training and testing set can be found in http://www.hlt.utdallas.edu/~vince/data/emnlp12/

References

Hector J. Levesque. 2011. *The Winograd Schema Challenge*. In AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum

Nathanael Chambers and Dan Jurafsky. 2008. *Unsupervised learning of narrative event chains*. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 787–797.

Altaf Rahman and Vincent Ng. 2012. *Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge*. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, page 777-789.

T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean. 2013. *Distributed Representations of Words and Phrases and their Compositionality*. In Proceedings of the Advanced Neural Information Processing Systems, page 3111-3119.

[2] The training and testing set can be found in http://www.hlt.utdallas.edu/~vince/data/emnlp12/