Cheap Facts and Counter-Facts

Rui Wang
Computational Linguistics Department
Saarland University
Room 2.04, Building C 7.4
Saarbruecken, 66123 Germany
rwang@coli.uni-sb.de

Chris Callison-Burch
Computer Science Department
Johns Hopkins University
3400 N. Charles Street (CSEB 226-B)
Baltimore, MD 21218, USA
ccb@cs.jhu.com.edu

Abstract

This paper describes our experiments of using Amazon's Mechanical Turk to generate (counter-)facts from texts for certain named-entities. We give the human annotators a paragraph of text and a highlighted named-entity. They will write down several (counter-)facts about this named-entity in that context. The analysis of the results is performed by comparing the acquired data with the recognizing textual entailment (RTE) challenge dataset.

1 Motivation

The task of RTE (Dagan et al., 2005) is to say whether a person would reasonably infer some short passage of text, the *Hypothesis* (H), given a longer passage, the *Text* (T). However, collections of such T-H pairs are rare to find and these resources are the key to solving the problem.

The datasets used in the RTE task were collected by extracting paragraphs of news text and manually constructing hypotheses. For the data collected from information extraction task, the H is usually a statement about a relation between two named-entities (NEs), which is written by expertise. Similarly, the H in question answering data is constructed using both the question and the (in)correct answers. Therefore, the research questions we could ask are,

1. Are these hypotheses really those ones people interested in?
2. Are hypotheses different if we construct them in other ways?
3. What would be a good negative hypotheses compared with the positive ones?

In this paper, we address these issues by using Amazon's Mechanical Turk (MTurk), online non-expert annotators (Snow et al., 2008). Instead of constructing the hypotheses targeted to IE or QA, we just ask the human annotators to come up with some facts they consider as relevant to the given text. For negative hypotheses, we change the instruction and ask them to write counter-factual but still relevant statements. In order to narrow down the content of the generated hypotheses, we give a focused named-entity (NE) for each text to guide the annotators.

2 Related Work

The early related research was done by Cooper et al. (1996), where they manually construct a textbook-style corpus aiming at different semantic phenomena involved in inference. However, the dataset is not large enough to train a robust machine-learning-based RTE system. The recent research from the RTE community focused on acquiring large quantities of textual entailment pairs from news headlines (Burger and Ferro, 2005) and negative examples from sequential sentences with transitional discourse connectives (Hickl et al., 2006). Although the quality of the data collected were quite good, most of the positive examples are similar to summarization and the negative examples are more like a comparison/contrast between two sentences instead of a contradiction. Those data are the real sentences used in news articles, but the way of obtaining them is not necessarily the (only) best way to

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 163–167,
Los Angeles, California, June 2010.

find entailment pairs. In this paper, we investigate an alternative inexpensive way of collecting entailment/contradiction text pairs by crowdsourcing.

In addition to the information given by the text, common knowledge is also allowed to be involved in the inference procedure. The Boeing-Princeton-ISI (BPI) textual entailment test suite[1] is specifically designed to look at entailment problems requiring world knowledge. We will also allow this in the design of our task.

3 Design of the Task

The basic idea of the task is to give the human annotators a paragraph of text with one highlighted named-entity and ask them to write some (counter-)facts about it. In particular, we first preprocess an existing RTE corpus using a named-entity recognizer to mark all the named-entities appearing in both T and H. When we show the texts to Turkers, we highlight one named-entity and give them one of these two sets of instructions:

Facts: Please write several facts about the highlighted words according to the paragraph. You may add additional common knowledge (e.g. Paris is in France), but please mainly use the information contained in the text. But please **do not copy and paste**!

Counter-Facts: Please write several statements that are contradictory to the text. Make your statements about the highlighted words. Please use the information mainly in the text. Avoid using words like **not** or **never**.

Then there are three blank lines given for the annotators to fill in facts or counter-factual statements. For each HIT, we gather facts or counter-facts for five texts, and for each text, we ask three annotators to perform the task. We give Turkers one example as a guide along with the instructions.

4 Experiments and Results

The texts we use in our experiments are the development set of the RTE-5 challenge (Bentivogli et al., 2009), and we preprocess the data using the Stanford named-entity recognizer (Finkel et al., 2005). In all, it contains 600 T-H pairs, and we use the texts to generate facts and counter-facts and hypotheses as references. We put our task online through CrowdFlower[2], and on average, we pay one cent for each (counter-)fact to the Turkers. CrowdFlower can help with finding trustful Turkers and the data were collected within a few hours.

	Total	Average (per Text)
Extracted NEs		
Facts	244	1.19
Counter-Facts	121	1.11
Generated Hypotheses		
Facts	790	3.85
Counter-Facts	203	1.86

Table 1: The statistics of the (valid) data we collect. The *Total* column presents the number of extracted NEs and generated hypotheses and the *Average* column shows the average numbers per text respectively.

To get a sense of the quality of the data we collect, we mainly focus on analyzing the following three aspects: 1) the statistics of the datasets themselves; 2) the comparison between the data we collect and the original RTE dataset; and 3) the comparison between the facts and the counter-facts.

Table 1 show some basic statistics of the data we collect. After excluding invalid and trivial ones[3], we acquire 790 facts and 203 counter-facts. In general, the counter-facts seem to be more difficult to obtain than the facts, since both the total number and the average number of the counter-facts are less than those of the facts. Notice that the NEs are not many since they have to appear in both T and H.

The comparison between our data and the original RTE data is shown in Table 2. The average length of the generated hypotheses is longer than the original hypotheses, for both the facts and the counter-facts. Counter-facts seem to be more verbose, since additional (contradictory) information is added. For instance, example ID 425 in Table 4, Counter_Fact_1 can be viewed as the more informative but contradictory version of Fact_1 (and the original hypoth-

[1] http://www.cs.utexas.edu/~pclark/bpi-test-suite/

[2] http://crowdflower.com/

[3] Invalid data include empty string or single words; and the trivial ones are those sentences directly copied from the texts.

esis). The average bag-of-words similarity scores are calculated by dividing the number of overlapping words of T and H by the total number of words in H. In the original RTE dataset, the entailed hypotheses have a higher BoW score than the contradictory ones; while in our data, facts have a lower score than the counter-facts. This might be caused by the greater variety of the facts than the counter-facts. Fact_1 of example ID 425 in Table 4 is almost the same as the original hypothesis, and Fact_2 of example ID 374 as well, though the latter has some slight differences which make the answer different from the original one. The NE position in the sentence is another aspect to look at. We find that people tend to put the NEs at the beginning of the sentences more than other positions, while in the RTE datasets, NEs appear in the middle more frequently.

In order to get a feeling of the quality of the data, we randomly sampled 50 generated facts and counter-facts and manually compared them with the original hypotheses. Table 3 shows that generated facts are easier for the systems to recognize, and the counter-facts have the same difficulty on average.

Although it is subjective to evaluate the *difficulty* of the data by human reading, in general, we follow the criteria that

1. Abstraction is more difficult than extraction;
2. Inference is more difficult than the direct entailment;
3. The more sentences in T are involved, the more difficult that T-H pair is.

Therefore, we view the Counter_Fact_1 in example ID 16 in Table 4 is more difficult than the original hypothesis, since it requires more inference than the direct fact validation. However, in example ID 374, Fact_1 is easier to be verified than the original hypothesis, and same as those facts in example ID 506. Similar hypotheses (e.g. Fact_1 in example ID 425 and the original hypothesis) are treated as being at the same level of difficulty.

After the quantitive analysis, let's take a closer look at the examples in Table 4. The facts are usually constructed by rephrasing some parts of the text (e.g. in ID 425, "after a brief inspection" is paraphrased by "investigated by" in Fact_2) or making a short summary (e.g. Fact_1 in ID 374, "George Stranahan spoke of Thompson's death."). For counter-facts, removing the negation words or changing into another adjective is one common choice, e.g. in ID 374, Counter_Fact_1 removed "n't" and Counter_Fact_3 changed "never" into "fully". The antonyms can also make the contradiction, as "rotten" to "great" in Counter_Fact_2 in ID 374.

Example ID 506 in Table 4 is another interesting case. There are many facts about Yemen, but no valid counter-facts are generated. Furthermore, if we compare the generated facts with the original hypothesis, we find that people tend to give straightforward facts instead of abstracts[4].

At last, we show some preliminary results on testing a baseline RTE system on this dataset. For the sake of comparison, we extract a subset of the dataset, which is balanced on entailment and contradiction text pairs, and compare the results with the same system on the original RTE-5 dataset. The baseline system uses a simple BoW-based similarity measurement between T and H (Bentivogli et al., 2009) and the results are shown in Table 5.

The results indicate that our data are slightly "easier" than the original RTE-5 dataset, which is consistent with our human evaluation on the sampled data

	Valid	Harder	Easier	Same
Facts	76%	16%	24%	36%
Counter-Facts	84%	36%	36%	12%

Table 3: The comparison of the generated (counter-)facts with the original hypotheses. The *Valid* column shows the percentage of the valid (counter-)facts; and other columns present the distribution of harder, easier cases than the original hypotheses or with the same difficulty.

	RTE-5	Our Data
Counter-/Facts	300/300	178/178
All "YES"	50%	50%
BoW Baseline	57.5%	58.4%

Table 5: The results of baseline RTE systems on the data we collected, compared with the original RTE-5 dataset. The *Counter-/Facts* row shows the number of the T-H pairs contained in the dataset; and the other scores in percentage are accuracy of the systems.

[4]But this might also be caused by the design of our task.

	Ave. Length	Ave. BoW	NE Position		
			Head	Middle	Tail
Original Entailment Hypotheses	7.6	0.76	46%	53%	1%
Facts	9.8	0.68	68%	29%	3%
Original Contradiction Hypotheses	7.5	0.72	44%	56%	0%
Counter-Facts	12.3	0.75	59%	38%	3%

Table 2: The comparison between the generated (counter-)facts and the original hypotheses from the RTE dataset. The *Ave. Length* column represents the average number of words in each hypothesis; The *Ave. BoW* shows the average bag-of-words similarity compared with the text. The three columns on the right are all about the position of the NE appearing in the sentence, how likely it is at the head, middle, or tail of the sentence.

(Table 3). However, it is still too early to draw conclusions based on the simple baseline results.

5 Conclusion and Future Work

In this paper, we report our experience of using MTurk to collect facts and counter-facts about the given NEs and texts. We find that the generated hypotheses are not entirely the same as the original hypotheses in the RTE data. One direct extension would be to use more than one NE at one time, but it may also cause problems, if those NEs do not have any relations in-between. Another line of research would be to test this generated resources using some real existing RTE systems and compare the results with the original RTE datasets, and also further explore the potential application of this resource.

Acknowledgments

The first author is supported by the PIRE scholarship program. The second author is supported by the EuroMatrixPlusProject (funded by the European Commission), by the DARPA GALE program under Contract No. HR0011-06-2-0001, and by the NSF under grant IIS-0713448. The views and findings are the authors' alone.

References

L. Bentivogli, B. Magnini, I. Dagan, H.T. Dang, and D. Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *Proceedings of the Text Analysis Conference (TAC 2009) Workshop*, Gaithersburg, Maryland, USA, November. National Institute of Standards and Technology.

John Burger and Lisa Ferro. 2005. Generating an entailment corpus from news headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 49–54, Ann Arbor, Michigan, USA. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. Using the framework. Technical Report LRE 62-051 D-16, The FraCaS Consortium.

I. Dagan, O. Glickman, and B. Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370. Association for Computational Linguistics.

Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing textual entailment with lcc's groundhog system. In *Proceedings of the Second PASCAL Challenges Workshop*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? Evaluating non- expert annotations for natural language tasks. In *Proceedings of EMNLP*.

ID: 16	Answer: **Contradiction**
Original Text	*The father of an Oxnard teenager accused of gunning down a gay classmate who was romantically attracted to him has been found dead, Ventura County authorities said today. Bill McInerney, 45, was found shortly before 8 a.m. in the living room of his Silver Strand home by a friend, said James Baroni, Ventura County's chief deputy medical examiner. The friend was supposed to drive him to a court hearing in his son's murder trial, Baroni said. McInerney's 15-year-old son, Brandon, is accused of murder and a hate crime in the Feb. 12, 2008, shooting death of classmate Lawrence "Larry" King, 15. The two boys had been sparring in the days before the killing, allegedly because Larry had expressed a romantic interest in Brandon.*
Original Hypothesis	*Bill McInerney is accused of killing a gay teenager.*
NE_1: Bill McInerney	
Counter_Fact_1	*Bill McInerney is still alive.*
ID: 374	Answer: **Contradiction**
Original Text	*Other friends were not surprised at his death. "I wasn't surprised," said **George Stranahan**, a former owner of the **Woody Creek Tavern**, a favourite haunt of Thompson. "I never expected Hunter to die in a hospital bed with tubes coming out of him." Neighbours have said how his broken leg had prevented him from leaving his house as often as he had liked to. One neighbour and long-standing friend, Mike Cleverly, said Thompson was clearly hobbled by the broken leg. "Medically speaking, he's had a rotten year."*
Original Hypothesis	*The **Woody Creek Tavern** is owned by **George Stranahan**.*
NE_1: George Stranahan	
Fact_1	*George Stranahan spoke of Thompson's death.*
Fact_2	*George Stranahan once owned the Woody Creek Tavern.*
Counter_Fact_1	*George Stranahan was surprised by his friend's death.*
Counter_Fact_2	*Medically, George Stranahan's friend, Humter Thompson, had a great year.*
Counter_Fact_3	*George Stranahan fully expected Thompson to die in a hospital with tubes coming out of him.*
NE_2: Woody Creek Tavern	
Fact_1	*Woody Creek Tavern was previously owned by George Stranahan.*
ID: 425	Answer: **Entailment**
Original Text	*Merseyside Police concluded after a brief inspection that the controversial blog **Liverpool Evil Cabal** does not break criminal law. However the council officers continue to search for the editor. The blog has been blocked on computers controlled by Liverpool Direct Ltd, a company jointly owned by Liverpool City Council and British Telecom. The council's elected officials have denied ordering the block and are currently investigating its origin.*
Original Hypothesis	***Liverpool Evil Cabal** is the name of an online blog.*
NE_1: Liverpool Evil Cabal	
Fact_1	*Liverpool Evil Cabal is a web blog.*
Fact_2	*Liverpool Evil Cabal was a blog investigated by the Merseyside Police.*
Counter_Fact_1	*Liverpool Evil Cabal is a blog of Liverpool Direct Ltd.*
Counter_Fact_2	*Liverpool Evil Cabal is freed from the charges of law breaking.*
ID: 506	Answer: **Entailment**
Original Text	*At least 58 people are now dead as a result of the recent flooding in **Yemen**, and at least 20,000 in the country have no access to shelter. Five people are also reported missing. The Yemeni government has pledged to send tents to help the homeless. The flooding is caused by the recent heavy rain in Yemen, which came as a shock due to the fact that the country only receives several centimeters of rain per year.*
Original Hypothesis	*Heavy rain caused flooding in **Yemen**.*
NE_1: Yemen	
Fact_1	*58 people are dead in Yemen because of flooding.*
Fact_2	*5 people in Yemen are missing.*
Fact_3	*At least 58 people are dead in Yemen because of flooding.*

Table 4: Examples of facts and counter-facts, compared with the original texts and hypotheses. We ask the Turkers to write several (counter-)facts about the highlighted NEs, and only part of the results are shown here.