A Dataset for Arabic Textual Entailment

Maytham Alabbas
School of Computer Science, University of Manchester, Manchester, M13 9PL, UK
alabbasm@cs.man.ac.uk
Department of Computer Science, College of Science, Basrah University, Basrah, Iraq
maytham.alabbas@gmail.com

Abstract

There are fewer resources for textual entailment (TE) for Arabic than for other languages, and the manpower for constructing such a resource is hard to come by. We describe here a semi-automatic technique for creating a first dataset for TE systems for Arabic using an extension of the 'headline-lead paragraph' technique. We also sketch the difficulties inherent in volunteer annotators-based judgment, and describe a regime to ameliorate some of these.

1 Introduction

One key task for natural language systems is to determine whether one natural language sentence entails another. One of the most popular generic tasks nowadays is called *textual entailment* (TE). Dagan and Glickman (2004) describe that *text T* textually entails *hypothesis H* if the truth of *H*, as interpreted by a typical language user, can be inferred from the meaning of *T*. For instance, (1a) entails (1b) whereas the reverse does not.

(1) a. *The couple are divorced.*
b. *The couple were married.*

Tackling this task will open the door to applications of these ideas in many areas of natural language processing (NLP), such as question answering (QA), semantic search, information extraction (IE), and multi-document summarisation.

Our main goal is to develop a TE system for Arabic. To achieve this goal we need firstly to create an appropriate dataset because there are, to the best of our knowledge, no such datasets available.

The remainder of this paper is organised as follows. The current technique for creating a textual entailment dataset is explained in Section 2 . Section 3 describes the Arabic dataset. A spammer detection technique is described in Section 4. Section 5 presents a summary discussion.

2 Dataset Creation

In order to train and test a TE system for Arabic, we need an appropriate dataset. We did not want to produce a set of *T-H* pairs by hand–partly because doing so is a lengthy and tedious process, but more importantly because hand-coded datasets are liable to embody biases introduced by the developer. If the dataset is used for training the system, then the rules that are extracted will be little more than an unfolding of information explicitly supplied by the developers. If it is used for testing then it will only test the examples that the developers have chosen, which are likely to be biased, albeit unwittingly, towards the way they think about the problem.

Our current technique for building an Arabic dataset for the TE task consists of two tools. The first tool is responsible for automatically collecting *T-H* pairs from news websites (Section 2.1), while the second tool is an online annotation system that allows annotators to annotate our collected pairs manually (Section 2.2).

2.1 Collecting *T-H* Pairs

A number of TE datasets have been produced for different languages, such as English,[1] Greek (Marzelou et al., 2008), Italian (Bos et al., 2009), German and Hindi (Faruqui and Padó, 2011). Some of these datasets were collected by the so-called *headline-lead paragraph* technique (Bayer et al., 2005; Burger and Ferro, 2005) from newspaper corpora, pairing the first paragraph of an article, as *T*, with its headline, as *H*. This is based on the observation that a news article's headline is very often a partial paraphrase of the first para-

[1] Available at: http://www.nist.gov/tac/2011/RTE/index.html

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 7–13,
Hissar, Bulgaria, 9-11 September 2013.

Source	Headline (Hypothesis)	Lead paragraph (Text)	Result
CNN	Berlusconi says he will not seek another term.	Italian Prime Minister Silvio Berlusconi said Friday he will not run again when his term expires in 2013.	YES
BBC	Silvio Berlusconi vows not to run for new term in 2013.	Italian Prime Minister Silvio Berlusconi has confirmed that he will not run for office again when his current term expires in 2013.	YES
Reuters	Berlusconi says he will not seek new term.	Italian Prime Minister Silvio Berlusconi declared on Friday he would not run again when his term expires in 2013.	YES

Figure 1: Some English *T-H* pairs collected by headline-lead paragraph technique.

graph of this article, conveying thus a comparable meaning.

We are building a corpus of *T-H* pairs by using headlines that have been automatically acquired from Arabic newspapers' and TV channels' websites[2] as queries to be input to Google via the standard Google-API. Then, we select the first paragraph, which usually represents the most related sentence(s) in the article with the headline (Bayer et al., 2005; Burger and Ferro, 2005), of each of the first 10 returned pages. This technique produces a large number of *T-H* pairs without any bias in either *Ts* or *Hs*. To improve the quality of the sentence pairs that resulted from the query, we use two conditions to filter the results: (i) the length of a headline must be at least more than five words to avoid very small headlines; and (ii) the number of common words (either in surface forms or lemma forms) between both sentences must be less than 80% of the headline length to avoid having excessively similar sentences. In the current work, we apply both conditions above to 85% of the *T-H* pairs from both training and testing sets. We then apply the first condition only to the remaining 15% of *T-H* pairs in order to leave some similar pairs, especially non entailments, to foil simplistic approaches (e.g. bag-of-words).

The problem here is that the headline and the lead-paragraph are often so similar that there would be very little to learn from them if they were used in the training phase of a TE system; and they would be almost worthless as a test pair–virtually any TE system will get this pair right, so they will not serve as a discriminatory test pair. In order to overcome this problem, we matched headlines from one source with stories from another. Using a headline from one source and the first sentence from an article about the same story but from another source is likely to produce *T-H* pairs which are not unduly similar. Figure 1 shows, for instance, the results of headlines from various sites (CNN, BBC and Reuters) that mention Berlusconi in their headlines on a single day.

We can therefore match a headline of one newspaper with related sentences from another one. We have tested this technique on different languages, such as English, Spanish, German, Turkish, Bulgarian, Persian and French. We carried out a series of informal experiment with native speakers and the results were encouraging, to the point where we took this as the basic method for suggesting *T-H* pairs.

Most of the Arabic articles that are returned by this process typically contain very long sentences (100+ words), where only a small part has a direct relationship to the query. With very long sentences of this kind, it commonly happens that only the first part of *T* is relevant to *H*. This is typical of Arabic text, which is often written with very little punctuation, with elements of the text linked by conjunctions rather than being broken into implicit segments by punctuation marks such as full stops and question marks. Thus what we really want as the text is actually the first conjunct of the first sentence, rather than the whole of the first sentence.

In order to overcome this problem, we simply need to find the first conjunction that links two sentences, rather than linking two substructures (e.g. two noun phrases (NPs)). MSTParser (McDonald and Pereira, 2006) does this quite reliably, so that parsing and looking for the first conjunct is a more reliable way of segmenting long Arabic sentences than simply segmenting the text at the first conjunction. For instance, selecting the second conjunction in segment (2) will give us the complete sentence *'John and Mary go to school in the morning'*, since it links two sentences. In contrast, selecting the first conjunction in segment (2) will give us solely the proper noun *'John'*, since it links two NPs (i.e. *'John'* and *'Mary'*).

[2]We use here Al Jazeera `http://www.aljazeera.net/`, Al Arabiya `http://www.alarabiya.net/` and BBC Arabic `http://www.bbc.co.uk/arabic/` websites as resources for our headlines.

(2) *John **and** Mary go to school in the morning **and** their mother prepares the lunch.*

2.2 Annotating *T-H* Pairs

The annotation is performed by volunteers, and we have to rely on their goodwill both in terms of how many examples they are prepared to annotate and how carefully they do the job. We therefore have to make the task as easy possible, to encourage them to do large numbers of cases, and we have to manage the problems that arise from having a mixture of people, with different backgrounds, as annotators. In one way having non-experts is very positive: as noted above, TE is about the judgements that a typical speaker would make. Not the judgements that a logician would make, or the judgements that a carefully briefed annotator would make, but the judgements that a typical speaker would make. From this point of view, having a mixture of volunteers carrying out the task is a good thing: their judgements will indeed be those of a typical speaker.

At the same time, there are problems associated with this strategy. Our volunteers may just have misunderstood what we want them to do, or they may know what we want but be careless about how they carry it out. We therefore have to be able to detect annotators who, for whatever reason, have not done the job properly (Section 4).

Because our annotators are geographically distributed, we have developed an online annotation system. The system presents the annotator with sentences that they have not yet seen and that are not fully annotated (here, annotated by three annotators) and asks them to mark this pair as positive 'YES', negative 'NO' and unknown 'UN'. The system also provides other options, such as revisiting a pair that they have previously annotated, reporting sentences that have such gross misspellings or syntactic anomalies that it is impossible to classify, skipping the current pair when a user chooses not to annotate this pair, and general comments (to send any suggestion about improving the system). The final annotation of each pair is computed when it is fully annotated by three annotators–when an annotator clicks 'Next', they are given the next sentence that has not yet been fully annotated. This has the side-effect of mixing up annotators: since annotators do their work incrementally, it is very unlikely that three people will all click 'Next' in lock-step, so there will be inevitable shuffling of annotators, with each person having a range of different co-annotators. All information about articles, annotators, annotations and other information such as comments is stored in a MySQL database.

3 Arabic TE Dataset

The preliminary dataset, namely Arabic TE dataset (ArbTEDS), consists of 618 *T-H* pairs. These pairs are randomly chosen from thousands of pairs collected by using the tool explained in Section 2.1. These pairs cover a number of subjects such as politics, business, sport and general news. We used eight expert and non-expert volunteer annotators[3] to identify the different pairs as 'YES', 'NO' and 'UN' pairs. Those annotators follow nearly the same annotation guidelines as those for building the RTE task dataset (Dagan et al., 2006). They used the online system explained in Section 2.2 to annotate our collected *T-H* pairs.

Table 1 summarises these individual results: the rates on the cases where an annotator agrees with at least one co-annotator (average around 91% between annotators) are considerably higher than those in the case where the annotator agrees with both the others (average around 78% between annotators). This suggests that the annotators found this is a difficult task. This table shows that comparatively few of the disagreements involve one or more of the annotators saying 'UN'–for 600 of the 618 pairs at least two annotators both chose 'YES' or both chose 'NO' (the missing 18 pairs arise entirely from cases where two or three annotators chose 'UN' or where one said 'YES', one said 'NO' and one said 'UN'. These 18 pairs are annotated as 'UN' and they are eliminated from our dataset, leaving 600 binary annotated pairs).

Agreement	YES	NO
$\geq$ 2 agree	478 (80%)	122 (20%)
3 agree	409 (68%)	69 (12%)

Table 1: ArbTEDS annotation rates.

As can be seen in Table 1, if we take the majority verdict of the annotators we find that 80% of the dataset are marked as entailed pairs, 20% as not entailed pairs. When we require unanimity between annotators, this becomes 68% entailed and

[3]All our annotators are Arabic native speaker PhD students, who are the author's colleagues. Some of them are linguistics students, whereas the others are working in fields related to NLP.

12% not entailed pairs. This drop in coverage, together with the fact that the ratio of entailed:not entailed moves from 100:25 to 100:17, suggests that relying on the majority verdict is unreliable, and we therefore intend to use only cases where all three annotators agree for both training and testing.

One obvious candidate is sentence length. It seems plausible that people will find long sentences harder to understand than short ones, and that there will be more disagreement about sentences that are hard to understand than about easy ones. Further statistical analysis results for the version of the dataset when there is unanimity between annotators are summarised in Table 2. We analyse the rates of this strategy that are shown in Table 1 according to the text's length, when the *H* average length is around 10 words and the average of common words between *T* and *H* is around 4 words. The average length of sentence in this dataset is 25 words per sentence, with some sentences containing 40+ words.

T's length	**#pairs**	**#YES**	**#NO**	**At least one disagree**
<20	131	97	11	23
20-29	346	233	38	75
30-39	110	69	20	21
>39	13	10	0	3
Total	**600**	**409**	**69**	**122**

Table 2: T's range annotation rates, three annotators agree.

Contrary to the expectation above, there does not seem to be any variation in agreement amongst annotators as sentence length changes. We therefore select the candidate *T-H* pairs without any restrictions on the length of the text to diversify the level of the examples' complexity, and hence to make the best use for our dataset.

3.1 Testing Dataset

It is worth noting in Table 1 that a substantial majority of pairs are marked positively–that *T* does indeed entail *H*. This is problematic, at least when we come to use the dataset for testing. For testing we need a balanced set: if we use a test set where 80% of cases are positive then a system which simply marks every pair positively will score 80%. It is hard, however, to get pairs where *T* and *H* are related but *T* does not entail *H* automatically. To solve this problem, we select the paragraph (other than the lead paragraph) in the article that shares the highest number of words with the headline for the first 10 returned pages. We called this technique *headline keywords-rest paragraph*. It produces a large number of potential texts, which are related to the main keywords of the headlines, without any bias.

In the case of testing set, we need a balanced 'YES' and 'NO' pairs (i.e. 50% pairs for each group). For this reason, we are currently following two stages to create our testset: (i) we apply our updated headline-lead paragraph technique for collecting positive pairs, since such technique is promising in this regard (see Table 1); and (ii) apply the strategy *headline keywords-rest paragraph* for collecting negative pairs and we will ask our annotators to select a potential text for each headline that it does not entail. Again we avoid asking the annotators to generate texts, in order to avoid introducing any unconscious bias. All the texts and hypotheses in our dataset were obtained from the news sources–the annotators' sole task is to judge entailment relations.

The preliminary results for collecting such dataset are promising. For instance, (3) shows example of positive pair where the annotators all agree for illustration.

(3) Positive pair

a. وزارة الدفاع الأمريكية البنتاغون تعد استراتيجية جديدة تضع الهجمات الألكترونية في مصاف

الأعمال الحربية حسبما ذكرت صحف أمريكية

wzArħ Al+dfAς Al+Âmrykyħ Al+bntAγwn tςd AstrAtyjyħ jdydħ tDς Al+hjmAt Al+Âlktrwnyħ fy mSAf Al+ÂςmAl Al+Hrbyħ HsbmA ðkrt SHf Âmrykyħ

"The US Department of Defense, the Pentagon, draw up a new strategy that categorises cyber-attacks as acts of war, according to US newspapers"

b. البنتاغون يعتبر الهجمات الألكترونية أعمالا حربية

Al+bntAγwn yςtbr Al+hjmAt Al+Âlktrwnyħ ÂςmAl Hrbyħ

"The Pentagon considers cyber-attacks as acts of war"

By applying the headline keywords-rest para-

graph on the entailed pair in (3), you could get not entailed pair as illustrated in (4).

(4) Negative pair for positive pair in (3)

a. صرح المتحدث باسم البنتاغون بأن: الرد على أي هجوم الكتروني تتعرض له الولايات المتحدة ليس ضروريا أن يكون بالمثل ولكن كل الخيارات مطروحة على الطاولة للرد على هذا الهجوم
SrH Al+mtHdθ bAsm Al+bntAγwn bÂn: Al+rd ς lý Ây hjwm Alktrwny ttςrD lh Al+wlAyAt Al+mtHdħ lys DrwryA Ân ykwn bAlmθl wlkn kl Al+xyarAt mTrwHħ llrd ς lý hðA Al+hjwm
"The Pentagon spokesman declared that: a response to any cyber-attacks on the US would not necessarily be a cyber-response and all options would be on the table to respond to this attack"

b. البنتاغون يعتبر الهجمات الألكترونية أعمالا حربية
Al+bntAγwn yςtbr Al+hjmAt Al+Âlktrwnyħ ÂςmAl Hrbyħ
"The Pentagon considers cyber-attacks as acts of war"

4 Spammer Checker

In order to check the reliability of our annotators, we used a statistical measure for assessing the reliability of agreement among our annotators when assigning categorical ratings to a number of annotating *T-H* pair of sentences. This measure is called *kappa*, which takes chance agreement into consideration. We use Fleiss's kappa (Fleiss, 1971), which is a generalisation of Cohen's kappa (Cohen, 1960) statistic to provide a measurement of agreement among a constant number of raters.

In our case, we need a global measure of agreement, which corresponds to the annotator reliability. We carry out the following steps:

1. The current annotator is ANT_i, $i=1$.
2. Create table for the ANT_i. This table includes all sentences annotated by ANT_i, and includes also as columns the other annotators who annotated the same sentences as ANT_i since each annotator has a range of different co-annotators. If an annotator does not annotate a sentence, then the corresponding cell should be left blank.
3. Compute the multiple-annotator version of kappa for all annotators in that table.
4. Compute another kappa for all annotators except ANT_i in that table.
5. If the kappa calculated in the step 4 exceeds that of step 3 significantly, then ANT_i is possibly a *spammer*.
6. $i=i+1$
7. If i exceeds 8 (i.e. number of our annotators), then stop.
8. Repeat this process from step 2 for the ANT_i.

To identify a 'spammer', you need to compare each annotator to something else (or some other group of annotators). If you take one annotator at a time, you will not be able to compute kappa, which takes chance agreement into consideration. You need two annotators or more to compute kappa.

We find out the kappa for each annotator with his/her co-annotators and another kappa for his/her co-annotators only for our eight annotators using the above steps, as shown in Table 3.

Annotator ID	**Kappa for current annotator**	**Kappa for co-annotators**
ANT_1	0.62	0.55
ANT_2	0.47	0.50
ANT_3	0.60	0.53
ANT_4	0.49	0.52
ANT_5	0.58	0.61
ANT_6	0.59	0.61
ANT_7	0.65	0.68
ANT_8	0.58	0.57
Average	**0.57**	**0.57**

Table 3: Reliability measure of our annotators.

The first thing to note about the results in Table 3 is that all kappa values between 0.4-0.79 represent a moderate to substantial level of agreement beyond chance alone according to the kappa interpretation given by Landis and Koch (1977) and Altman (1991). Also, the variation between the kappa including an annotator and the kappa of his/her co-annotators only is comparatively slight for all annotators. The average of both kappas for all annotators is equal (i.e. 0.57), which suggests

that the strength of agreement among our annotators is moderate (i.e. $0.4 \leq kappa \leq 0.59$). We have solely three annotators (ANT_1, ANT_3 and ANT_8) where the kappas including them are higher than kappas for their co-annotators. The other annotators have kappas less than the kappas of their co-annotators but these differences are very slight. These findings suggest that all our annotators are reasonably reliable and we can use their annotated dataset in our work, but they also provide us with an indication of who is most reliable for tasks such as the extra annotation described in Section 3.1.

5 Summary

We have outlined an approach to the task of creating a first dataset for a TE task for working with a language where we have to rely on volunteer annotators. To achieve this goal, we tested two main tools. The first tool, which depends on the Google-API, is responsible for acquisition of *T-H* pairs based on the headline-lead paragraph technique of news articles. We have updated this idea in two ways: (i) for training dataset, we use the lead paragraph from an article with a closely linked headline. This notion is applicable to the collection of such a dataset for any language. It has two benefits. Firstly, it makes it less likely that the headline will be extracted directly from the sentence that it is being linked to, since different sources will report the same event slightly differently. Secondly, it will be more likely than the original technique to produce *T-H* pairs where *T* entails *H* with few common words between *T* and *H*; and (ii) for testing dataset, we use the same technique for training except that we take the paragraph from the rest of the article (i.e. each paragraph in the article except the lead one) that gives the highest number of common words between both headline and paragraph. This is particularly important for testing, since for testing you want a collection which is balanced between pairs where *T* does entail *H* and ones where it does not. This technique will be more likely than the original technique and the updated technique for training to produce *T-H* pairs where *T* does not entail *H* with partly higher common words between *T* and *H*, which will pose a problem to a TE system. Automatically obtaining *T-H* pairs where *T* is reasonably closely linked to *H* but does not entail it is quite tricky. If the two are clearly distinct then they will not pose a very difficult test. As shown in Table 1, by using updated headline-lead paragraph technique, we have a preponderance of positive examples, but there is a non-trivial set of negative ones, so it is at least possible to extract a balanced test set. We therefore apply the headline keywords-rest paragraph technique to construct a balanced test set from our annotated dataset.

In order to make sure that our data is reliable, we check unreliable annotator(s) using kappa coefficient based strategy, which takes chance into consideration rather than agreement between annotators only. This strategy suggests that all our annotators are reliable.

We intend to make our dataset available to the scientific community thus allowing other researchers to duplicate their methodology and confront the results obtained.

Acknowledgments

My special thanks and appreciations go to my supervisor, Professor Allan Ramsay (UK), Dr. Kilem L. Gwet (USA) and Dr. Yasser Sabtan (Egypt) for productive discussions. I also owe my deepest gratitude to Iraqi Ministry of Higher Education and Scientific Research for financial support in my PhD study.

References

Douglas G. Altman. 1991. *Practical Statistics for Medical Research*. Chapman and Hall, London, UK.

Samuel Bayer, John Burger, Lisa Ferro, John Henderson, and Alexander Yeh. 2005. MITRE's submissions to the EU PASCAL RTE Challenge. In *Proceedings of the1st PASCAL Recognising Textual Entailment Challenge*, pages 41–44, Southampton, UK.

Johan Bos, Fabio Massimo Zanzotto, and Marco Pennacchiotti. 2009. Textual entailment at EVALITA 2009. In *Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, pages 1–7, Reggio Emilia, Italy.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: generic applied modeling of language variability. In *PASCAL Workshop on Learning Methods for Text Understanding and Mining*, pages 26–29, Grenoble, France.

Ido Dagan, Oren Glickman and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer Berlin-Heidelberg.

Manaal Faruqui and Sebastian Padó. 2011. Acquiring entailment pairs across languages and domains: a data analysis. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS'11)*, pages 95–104, Oxford, UK. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Evi Marzelou, Maria Zourari, Voula Giouli, and Stelios Piperidis. 2008. Building a Greek corpus for textual entailment. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pages 1680–1686, Marrakech, Morocco. European Language Resources Association.

Ryan McDonald and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 81–88, Trento, Italy. Association for Computational Linguistics.