A Corpus of General and Specific Sentences from News

Annie Louis, Ani Nenkova

University of Pennsylvania
Philadelphia, PA 19104
lannie@seas.upenn.edu, nenkova@seas.upenn.edu

Abstract

We present a corpus of sentences from news articles that are annotated as *general* or *specific*. We employed annotators on Amazon Mechanical Turk to mark sentences from three kinds of news articles—reports on events, finance news and science journalism. We introduce the resulting corpus, with focus on annotator agreement, proportion of general/specific sentences in the articles and results for automatic classification of the two sentence types.

Keywords: General-Specific, Mechanical Turk, News corpus

1. Introduction

Consider the two sentences below. While the first one conveys a topic it does not provide details. In contrast, the second supplies precise information. We call sentences like the first one "general" and sentences like the second "specific".

> [1] Now, the personal-computer revolution is finally reaching Japan.
>
> [2] While American PC sales have averaged roughly 25% annual growth since 1984 and West European sales a whopping 40%, Japanese sales were flat for most of that time.

Texts contain a mix of general and specific sentences and the distinction could be incorporated in many language applications such as essay grading, question answering and information retrieval. For example, a well-written essay can be expected to have several topic statements and each of these followed by specific arguments that support the general facts. There are also studies on the academic writing genre showing that conference papers have a structure where introductions and conclusions are general and details are presented in the sections in between (Swales and Feak, 1994). By identifying the amount and sequence of general and specific content we can evaluate the writing quality of such texts. Further, a question answering or information retrieval system can tailor answers depending on the detail needed. Some queries and users may require general information, others seek specific details.

In order to facilitate the development of automatic detectors for sentence specificity, we have created a corpus where judges marked individual sentences as general or specific.[1] This paper presents the details and results from the annotations. A key property of our corpus is its diversity. It contains sentences from three different types of news articles—event reports, finance and science journalism. We conducted the annotations through non-experts on Mechanical Turk who were provided with minimal instructions and training. From our results, we found that the distinction is fairly intuitive to people and the annotators could make reliable judgements on sentences from all three genres.

We have also developed an automatic method for predicting a sentence as general or specific which can make the distinction with close to 80% accuracy on the data we annotated.

2. Annotation Details

In this section, we describe the articles we have chosen for annotation and detail how the annotations were carried out on Mechanical Turk.

2.1. Data

Our corpus has news articles from three existing corpora which span different genres. We also list the document identifiers of the articles within each corpus.

AQUAINT: We chose 8 articles from the AQUAINT corpus (Graff, 2002) which is traditionally used for question answering and summarization. Six of them are news reports published by Associated Press and two are from Financial Times. Most articles here are short and we enforced a minimum length limit of 30 sentences. There are 292 sentences in the 8 articles combined. [docid: AP880713-0175, FT931-3664, AP900131-0200, FT923-5589, AP901019-0072, AP891116-0035, AP890922-0007, AP881002-0048]

WSJ: The Wall Street Journal corpus (Marcus et al., 1994) has mostly finance news articles. We chose three articles from the WSJ and these are longer than those from AQUAINT, each about 100 sentences. The set has a total of 294 sentences. [docid: wsj-0445, wsj-1037, wsj-1394]

NYT-science: We chose three articles reporting science news from the New York Times corpus (Sandhaus, 2008). While still news, these articles are quite different compared to the rest. For example, one of the articles discusses how the concentration of carbon dioxide in the atmosphere has changed over time. A total of 308 sentences were annotated from this source. [docid: 2002-03-05-1373005, 2006-11-07-1802956, 2007-05-10-1846387]

[1] The dataset can be downloaded from `http://www.cis.upenn.edu/~lannie/genspec.html`

Apart from difference in the topic and content, these articles also differ in writing style. The articles from AQUAINT are mostly event-oriented, reporting important facts around a particular current issue. In contrast, the NYT-science articles are descriptive and explanatory. The articles here can also take the form of narratives involving people and a storyline, and this was in fact the case for one of the articles we annotated.

2.2. Mechanical Turk Annotation

We provided the sentences to annotators on Amazon Mechanical Turk[2]. Each sentence was annotated by five different assessors. They marked a sentence as either general, specific or "cannot decide". We briefly described the difference between general and specific sentences and gave examples. The assessors largely relied on their intuition to mark the distinction. We provided the following instructions.

"Sentences could vary in how much detail they contain. One distinction we might make is whether a sentence is general or specific. General sentences are broad statements about a topic. Specific sentences contain details and can be used to support or explain the general sentences further. In other words, general sentences create expectations in the minds of a reader who would definitely need evidence or examples from the author. Specific sentences can stand by themselves. For example, one can think of the first sentence of an article or a paragraph as a general sentence compared to one which appears in the middle. In this task, use your intuition to rate the given sentence as general or specific.[3] Some examples are provided below but they do not cover all the sentence types you may encounter."

Examples: (These examples were taken from New York Times science section but are different from the articles given for annotation.)

GENERAL SENTENCES:

> [G1] A handful of serious attempts have been made to eliminate individual diseases from the world.
> [G2] In the last decade, tremendous strides have been made in the science and technology of fibre optic cables.
> [G3] Over the years interest in the economic benefits of medical tourism has been growing.

SPECIFIC SENTENCES:

> [S1] In 1909, the newly established Rockefeller Foundation launched the first global eradication campaign, an effort to end hookworm disease, in fifty-two countries.
> [S2] Solid silicon compounds are already familiar–as rocks, glass, gels, bricks, and of course, medical implants.
> [S3] Einstein undertook an experimental challenge that had stumped some of the most adept lab hands of all time–explaining the mechanism responsible for magnetism in iron.

Agree	**AQ**	**WSJ**	**NYT-science**
5	108	96	82
4	91	102	121
3	88	95	102
Undecided	5	1	3
Total	292	294	308

Table 1: The number of sentences for each agreement category. Agree 5 means all 5 annotators agreed on the class for a sentence.

	AQUAINT	
Agree	**General**	**Specific**
5	33 (28.2)	75 (44.1)
4	35 (29.9)	56 (32.9)
3	49 (41.8)	39 (22.9)
Total	117	170

	WSJ	
Agree	**General**	**Specific**
5	51 (31.8)	45 (33.8)
4	57 (35.6)	45 (33.8)
3	52 (32.5)	43 (32.3)
Total	160	133

	NYT-science	
Agree	**General**	**Specific**
5	32 (25.6)	50 (27.7)
4	48 (38.4)	73 (40.5)
3	45 (36.0)	57 (31.6)
Total	125	180

Table 2: The annotator agreement numbers split by type of majority class

3. Analysis of Annotations

In this section, we discuss the agreement between assessors and differences we observed among the three types of news sources that were annotated.

3.1. Annotator Agreement

On Mechanical Turk, we obtained annotations from five different assessors for each sentence. However, the same five assessors did not annotate all the sentences so we are not able to report the standard Kappa measures. Instead, we present the number of sentences split by how many annotators agreed on it (Table 1).

We find that across all three sets of sentences, about two-thirds (200 sentences) have an agreement level of 4 or 5. These agreement numbers are high given that annotators followed their intuition.

It is also informative to analyze the agreement numbers split by general/specific distinction. We wanted to know if agreement is higher for one of the sentence types. Table 2 reports the agreement per category for the three data sets.

[2] `http://sites.google.com/site/amtworkshop2010/`

[3] An option of selecting "cannot decide" was also given to the assessors.

	Agreement 5
General	[NYT] Climatologists and policy makers, they say, need to ponder such complexities rather than trying to ignore or dismiss the unexpected findings. [AQ] There are two standard explanations why a weak dollar prompts bond prices to fall. [WSJ] In the private sector, practically every major company is setting explicit goals to increase employees' exposure to computers.
Specific	[NYT] Isabella Bailey, Anya's mother, said she had no idea that children might be especially susceptible to Risperdal's side effects. [AQ] WAAY reported at least one person died when the roof of a business collapsed from winds that overturned cars in the area. [WSJ] Apple didn't introduce a kanji machine – one that handles the Chinese characters of written Japanese – until three years after entering the market.
	Agreement 3
General	[NYT] "The geologic record over the past 550 million years indicates a good correlation," said Robert A. Berner, a Yale geologist and pioneer of paleoclimate analysis. [AQ] He accomplished the same feat in 1980 and became the first man to sweep the events twice. [WSJ] As with many other goods, the American share of Japan's PC market is far below that in the rest of the world.
Specific	[NYT] In 2004, Dr. Berner of Yale and four colleagues fired back. [AQ] East Germany had 102 medals and 37 gold, and the United States 94 medals and 36 gold. [WSJ] "If it were an open market, we would have been in in 1983 or 1984," says Eckhard Pfeiffer, who heads Compaq Computer Corp.'s European and international operations.

Table 3: Example general and specific sentences with agreement 5 and 3

On NYT and WSJ sentences, the judges have similar agreement on examples from both general and specific class. But on the AQUAINT corpus, the agreement on the general sentences is lower than that on the other sets (58% at level 4 or 5) but the agreement is considerably better when the sentence is specific (77% have agreement of 4 or 5). So the specific sentences from the AQUAINT corpus appear to be easier for annotators. But on the whole, our judges made reliable judgements on both general and specific sentences.

3.2. Cases of Disagreement

In Table 3, we present example sentences with full agreement and those with low agreement from our three datasets. The sentences with lower agreement appear to exhibit a genuine mix of general and specific characteristics. For example the first specific sentence with agreement level 3 has details about the year of the event and the people involved but the event itself is not specified. Similarity the first general sentence with low agreement has detailed description of the geologist but the findings that he reports are fairly general. This evidence from the annotators indicates that the distinction between general and specific can be treated more transparently as a matter of degree rather than as fixed binary classes.

We also observe the influence of context. Since the sentences are annotated out of context, sometimes, the sentences can be interpreted as general because they have pronouns and other links which appear unspecified but would be easily clear given surrounding sentences. For example, in the second specific sentence with low agreement (in Table 3), details about which medals were won are reported but one does not know the sports event they are associated with. When this information is also presented, we can expect that annotators might rate this sentence as specific with much more agreement. In future annotations, we plan to have a dedicated class for this type of lack of specificity. Such extended distinctions would be helpful for summarization and question-answering systems which will obviously benefit from being able to identify sentences whose interpretation relies on context.

3.3. Distribution of General and Specific Sentences

We found that 40% of sentences from AQUAINT and NYT-science were marked as general and the value is 54.4% for WSJ. So while NYT and AQUAINT have more specific sentences than general, WSJ shows an opposite trend.

In future work, we wish to analyze in detail what is the right amount of general/specific content for different texts. Some of the factors could be the article topic, its length and writing style. When we studied automatically generated summaries, we found that people find the summaries to be of better content quality when they have a balance of general and specific content (Louis and Nenkova, 2011b).

4. Classification Experiments

We have also developed a classifier to predict general versus specific sentences and it is described in detail in Louis and Nenkova (2011a). Our features include sentence length, count of polarity words, adjectives and different types of syntactic phrases. We also include the presence of named entities and numbers, the likelihood under language models trained on news as well as the idf of words in the sentence. These features form our non-lexical set. We also include the presence of each word in the sentence as features and call them the lexical category.

4.1. Prediction Accuracies

In Table 4, we report the 10-fold cross validation accuracies of these features on all the data from our annotations combined (total of 885 sentences). We used the majority

Features	Accuracy
non-lexical	79.43
lexical	71.52
non-lexical + lexical	78.19

Table 4: Results for automatic prediction of general and specific sentences

Agree	Correct	Wrong
5	0.87 (4,3)	0.65
4	0.81 (3)	0.68
3	0.74	0.70

Table 5: The average confidence of the classifier for correct and wrong predictions. The examples are split across the agreement levels

judgement as the class of each sentence. Overall, there are slightly more specific sentences (55%) than general. So a random baseline that predicted the majority specific class would get 55% accuracy. We trained the classifier using logistic regression.

The non-lexical features perform best, with 79% accuracy. The word features are sparse given our small dataset but still give about 72% accuracy. The combination of the two categories was not helpful on this dataset.

Above we have presented the results where we trained our classifier using the annotations we have obtained. In Louis and Nenkova (2011a), we report results using an expanded training set of general and specific sentences. These additional sentences were obtained from existing annotations of certain types of discourse relations from the Penn Discourse Treebank (Prasad et al., 2008). The INSTANTIATION type discourse relation involves two sentences: the second provides an example for the fact presented in the first sentence. We used the first sentence in these relations as general and the second as specific (no pairing information was preserved) and obtained a much larger set of examples. Results using this larger training set are described in Louis and Nenkova (2011a). On this expanded collection, we found that the accuracy of the lexical feature set improves and both lexical and non-lexical categories give similar performance. Further, when we trained the classifier using only examples from the discourse relations and tested them on the manually annotated sentences, we obtained 75% accuracy showing that the examples from discourse relations are good exemplars for general-specific distinction. So other researchers can also use these relations to expand the annotations we have presented here.

4.2. Analysis of Classifier Confidence

We now turn back to the issue of assessor agreement and perform some analyses to understand how the classifier handles examples with different agreement levels. Specifically, we studied the relationship between the confidence from the classifier (logistic regression probability) and the annotator agreement on an example. We first combined the predictions for the sentences from the 10 folds and split the data into sentences which the classifier predicted correctly (above 0.5 confidence for the right class) and wrong predictions (above 0.5 confidence for the wrong class). Then in each set, we recorded the average value of classifier confidence on examples with different agreement. The results are shown in Table 5. When the mean value in one agreement level is significantly higher (under a two-sided t-test) than at another level, the lower levels are shown within parentheses.

We find that when the prediction is correct, the confidence on the examples with highest agreement is on average larger than that on lower agreement levels. On the wrong predictions, we see an opposite trend. On the examples where annotators agreed highly that they belong to one category, the classifier makes lower confidence predictions. On the lower agreement examples, it mispredicts with higher confidence indicating more confusion. Although for all the data combined, the values in the wrong prediction column are not statistically significant, when split by corpus, the AQUAINT data shows significant results for these numbers. This finding indicates that the classifier confidence varies according to the annotator agreement even though this information was not available to the classifier. This is an additional motivation to treat the general/specific distinction as a matter of degree rather than strict binary classes, and the classifier confidence values can be utilized as a measure of graded distinction.

5. Conclusion

We have presented the first corpus of sentence level general/specific distinction. Even though it is a new task, we found that non-expert annotators could make highly reliable judgements. There are avenues for improvement in both annotation scheme and the use of these distinctions in applications. We hope that our corpus will be beneficial to researchers for investigating this dimension further.

6. References

D. Graff. 2002. The aquaint corpus of english news text. *Corpus number LDC2002T31, Linguistic Data Consortium, Philadelphia*.

A. Louis and A. Nenkova. 2011a. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of IJCNLP*.

A. Louis and A. Nenkova. 2011b. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation, ACL-HLT*, pages 34–42.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC*.

E. Sandhaus. 2008. The new york times annotated corpus. *Corpus number LDC2008T19, Linguistic Data Consortium, Philadelphia*.

J. M. Swales and C. Feak. 1994. *Academic writing for graduate students: A course for non-native speakers of English*. Ann Arbor: University of Michigan Press.