Automatic identification of general and specific sentences by leveraging discourse annotations

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

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

Abstract

In this paper, we introduce the task of identifying general and specific sentences in news articles. Given the novelty of the task, we explore the feasibility of using existing annotations of discourse relations as training data for a general/specific classifier. The classifier relies on several classes of features that capture lexical and syntactic information, as well as word specificity and polarity. We also validate our results on sentences that were directly judged by multiple annotators to be general or specific. We analyze the annotator agreement on specificity judgements and study the strengths and robustness of features. We also provide a task-based evaluation of our classifier on general and specific summaries written by people. Here we show that the specificity levels predicted by our classifier correlates with the intuitive judgement of specificity employed by people for creating these summaries.

1 Introduction

Sentences in written text differ in how much specific content they have. Consider the sentences in Table 1 from a news article about the Booker prize. The first one is specific and details the issues surrounding the books chosen for the award. The second sentence is general, it states that the prize is controversial but provides no details. In this work, we present the first analysis of properties associated with general and specific sentences and introduce an approach to automatically identify the two types.

The distinction between general and specific sentences would be beneficial for several applications. Prescriptive books on writing advise that sentences that make use of vague and abstract words should be avoided or else immediately followed by specific clarifications (Alred et al., 2003). So our classifier could be useful for the prediction of writing quality. Other applications include text generation systems which should control the type of content produced and information extraction systems can use the distinction to extract different types of information.

The novel, a story of Scottish low-life narrated largely in Glaswegian dialect, is unlikely to prove a popular choice with booksellers who have damned all six books shortlisted for the prize as boring, elitist and - worst of all - unsaleable.
...
The Booker prize has, in its 26-year history, always provoked controversy.

Table 1: General (in italics) and specific sentences

Our definition of general/specific is based on the level of detail present in a sentence. This definition contrasts our work from some other recent studies around the idea of generic/specific distinctions in text. Reiter and Frank (2010) present an automatic approach to distinguish between noun phrases which describe a class of individuals (generic) versus those which refer to a specific individual(s). In Mathew and Katz (2009), the aim is to distinguish sentences which relate to a specific event (called episodic) from those which describe a general fact (habitual sentences). Our focus is on a different and broader notion of general/specific which is motivated by potential applications in summarization and writing feedback. The task of identifying these types of sentences has not been addressed in prior work.

We present a supervised classifier for detecting general and specific sentences. We obtain our training data from the Penn Discourse Treebank (PDTB), where relevant distinctions have been annotated in the larger context of discourse relation analysis. We show that classification accuracies as high as 75% can be obtained for distinguishing sentences of the two types compared with a ran-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 605–613,
Chiang Mai, Thailand, November 8 – 13, 2011.

dom baseline of 50%. We also perform an annotation study to obtain direct judgements from people about general/specific sentences in news articles from two corpora and use this dataset to validate the accuracy of our features and their robustness across genre. Finally, we train a classifier on the combined set of all annotated data.

We also present a task-based evaluation of our classifier using a large corpus of summaries written by people. For some of the topics, people were instructed to write specific summaries that focus on details, for others they were asked to include only general content. We find that our classifier successfully predicts the difference in specificity between these two types of summaries.

2 A general vs. specific sentence classifier based on discourse relations

The task of differentiating general and specific content has not been addressed in prior work, so there is no existing corpus annotated for specificity. For this reason, we first exploit indirect annotations of these distinctions in the form of certain types of discourse relations annotated in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008). The discourse relations we consider are Specification and Instantiation. They are defined to hold between adjacent sentences. The definitions of the relations do not talk directly about the specificity of sentences, but they seem to indirectly indicate that the first one is general and the second is specific. The exact definitions of these two relations in the PDTB are given in (Prasad et al., 2007). Some examples are shown in Table 2.

The PDTB annotations cover 1 million words from Wall Street Journal (WSJ) articles. Instantiations and Specifications are fairly frequent (1403 and 2370 respectively). In contrast to efforts in automatic discourse processing (Marcu and Echihabi, 2001; Sporleder and Lascarides, 2008), in our work we are not interested in identifying adjacent sentences between which this relation holds. Our idea is to use the *first* sentences in these relations as general sentences and the *second* as specific sentences.[1] Although the definitions of these relations describe the specificity of one sentence relative to the other, we do not focus on this pairwise difference in specificity. We believe that the realization of a general sentence should have some unique properties regardless of the particular sentence that precedes or follows it.

We use these relations to study the properties of general and specific sentences and to test the feasibility of differentiating these two types. We obtain good success on this task and equipped with the knowledge from our study, we collect direct judgements of general/specific notion from annotators on a smaller set of sentences. Using these annotations, we confirm that our classifier learnt on the discourse relations generalizes without noticeable compromise in accuracy. We describe our classifier based on discourse relations here, the annotation study is detailed in the next section.

2.1 Features

Based on a small development set of 10 examples each of Instantiation and Specification, we came up with several features that distinguished between the specific and general sentences in the sample. Some of our features require syntax information. We compute these using the manual parse annotations for the articles from the Penn Treebank corpus (Marcus et al., 1994).

Sentence length. We expected general sentences to be shorter than the specific ones. So we introduced two features—the number of words in the sentence and the number of nouns.

Polarity. Sentences with strong opinion are typical in the general category in our examples in Table 2. For instance, the phrases "publishing sensation", and "very slowly–if at all" are evaluative while the specific sentences in these relations present evidence which justify the general statements. So, we record for each sentence the number of positive, negative and polar (not neutral) words using two lexicons—The General Inquirer (Stone et al., 1966) and the MPQA Subjectivity Lexicon (Wilson et al., 2005). We also add another set of features where each of these counts is normalized by the sentence length.

Specificity. Specific sentences are more likely to contain specific words and details. We use two sets of features to capture specificity of words in the sentence. The first of these is based on WordNet (Miller et al., 1990) and is motivated by prior work by Resnik (1995) where hypernym relations from WordNet were used to compute specificity. For each noun and verb in a sentence, we record the length of the path from the word to the root of

[1] We use only the *implicit* relations from the PDTB; ie, the sentences are not linked by an explicit discourse connective such as 'because' or 'but' that signals the relation.

Instantiations

[1] *The 40-year-old Mr. Murakami is a publishing sensation in Japan.* A more recent novel, " Norwegian Wood " (every Japanese under 40 seems to be fluent in Beatles lyrics), has sold more than four million copies since Kodansha published it in 1987.

[2] *Sales figures of the test-prep materials aren't known, but their reach into schools is significant.* In Arizona, California, Florida, Louisiana, Maryland, New Jersey, South Carolina and Texas, educators say they are common classroom tools.

[3] *Despite recent declines in yields, investors continue to pour cash into money funds.* Assets of the 400 taxable funds grew by $ 1.5 billion during the last week, to $ 352.7 billion.

Specifications

[4] *By most measures, the nation's industrial sector is now growing very slowly—if at all.* Factory payrolls fell in September.

[5] *Mrs. Hills said that the U.S. is still concerned about 'disturbing developments in Turkey and continuing slow progress in Malaysia.'* She didn't elaborate, although earlier U.S. trade reports have complained of videocassette piracy in Malaysia and disregard for U.S. pharmaceutical patents in Turkey.

[6] *Alan Spoon, recently named Newsweek president said Newsweek's ad rates would increase 5% in January.* A full, four-color page in Newsweek will cost $100,980

Table 2: Examples of general (in italics) and specific sentences from the PDTB

the WordNet hierarchy through the hypernym relations. The longer this path, we would expect the words to be more specific. The average, min and max values of these distances are computed separately for nouns and verbs and are used as features.

Another measure of word specificity is the inverse document frequency (idf) for a word w (Joho and Sanderson, 2007), defined as $\log \frac{N}{n}$. Here N is the number of documents in a large collection, and n is the number of documents that contain the word w. We use articles from one year (87,052 documents) of the New York Times (NYT) corpus (Sandhaus, 2008) to compute idf. Words not seen in the NYT corpus were treated as if they were seen once. The features for a sentence are the average, min and max idfs for words in the sentence.

NE+CD. In news articles, especially the WSJ, specific sentences often contain numbers and dollar amounts. So we add as features the count of numbers (identified using the part of speech), proper names and dollar signs. The performance of these features, however, is likely to be genre-dependent. We also introduce another entity-related feature—the number of plural nouns. From our example sentences, we notice that plural quantities or sets are a property of general sentences.

Language models. General sentences often contain unexpected, catchy words or phrases. Consider the phrase "pour cash" in example [3] (Table 2); it is figurative and informal in the context of finance reports. When one reads the second sentence in the relation and observes the actual rise in funds investments, we understand why such a figurative phrase was used to introduce this fact. We expected that language models would capture this aspect by assigning a lower likelihood to unexpected content in the general sentences. We build unigram, bigram and trigram language models using one year of news articles from the NYT corpus. Using each model, we obtain the log probability and perplexity of the sentences to use as features. The unigram language model captures the familiarity of individual words. On the other hand, we expect the perplexity computed using higher order models to distinguish between common word transitions in the domain, and those that are unexpected and evoke surprise.

Syntax. We also noted frequent usage of qualitative words such as adjectives and adverbs in general sentences. So we include some syntax based features: counts of adjectives, adverbs, adjective phrases and adverbial phrases. We also record the number of verb phrases and their average length in words and the number of prepositional phrases. We expect that longer verb phrases would be associated with more specific sentences.

Words. We also add the count of each word in the sentence as a feature. Numbers and punctuations were removed but all other words were included. Only words seen in the training set are valid features. New words in the test sentences are ignored.

2.2 Results

We build two classifiers for distinguishing general and specific sentences: one trained on sentences from Instantiation relations, and one on sentences from Specification. The first sentence in the relation was considered an example of general sentence, and the second of specific one. No pairing information was preserved or exploited. We train a logistic regression classifier[2] with each set of features described above and evaluate the predictions

[2] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

Features	Instantiations	Specifications
NE+CD	68.6	56.1
language models	65.8	55.7
specificity	63.6	57.2
syntax	63.3	57.3
polarity	63.0	53.4
sentence length	54.0	57.2
all non-lexical	75.0	62.0
lexical (words)	74.8	59.1
all features	75.9	59.5

Table 3: Classifier accuracy (baseline 50%)

using 10-fold cross validation.

We choose logistic regression for our task because we expected that a probability measure would be more appropriate to associate with each sentence rather than hard classification into the two classes. We provide further analysis of the classifier confidence in the next section. Here for reporting results, we use a threshold value of 0.5 on the confidence score. There are equal number of positive and negative examples, so the baseline random accuracy is 50%. Table 3 shows the accuracy of our features.

The classifiers trained on Instantiation examples are promising and better than those trained on Specifications. The highest accuracy on Instantiations-based classifier comes from combining all features, reaching 75.9% which is more than 25% absolute improvement over the baseline. The individually best class of features are the words with 74.8% accuracy showing that there are strong lexical indicators of the distinction.

Among the non-lexical features, the NE+CD class is the strongest with an accuracy of 68%. Language models, syntax, polarity and specificity features are also good predictors, each outperforming the baseline by over 10% accuracy. The sentence length features are the least indicative. These non-lexical feature classes though not that strong individually, combine to give the same performance as the word features. Moreover, one would expect that non-lexical features would be more robust across different types of news and topics compared to the lexical ones and would have fewer issues related to data sparsity. We analyse this aspect in the next section.

For the Specifications-based classifier, the highest performance is barely 10% above baseline. The best accuracy (62%) is obtained with a combination of all non-lexical features. In contrast to the Instantiations case, language models and entities features sets are less accurate in making the general-specific distinction on the Specification examples. Polarity is the worst set of features with only 53% accuracy.

A possible explanation of the difference in results from the two types of training data is that in Specification relations, the specificity of the second sentence is only relative to that of the first. On the other hand, for Instantiations, there are individual characteristics related to the generality or specificity of sentences. We confirm this hypothesis in Section 3.2.

2.3 Feature analysis

In this section, we take a closer look at the features that most successfully distinguished specific and general sentences on the *Instantiation* dataset. Given that words were the most predictive feature class, we identified those with highest weight in the logistic regression model. Here we list the top word features for the two types of sentences and which appear in at least 25 training examples.

General number, but, also, however, officials, some, what, prices, made, lot, business, were

Specific one, a, to, co, i, called, we, could, get, and, first, inc

Discourse connectives such as 'but', 'also' and 'however', and vague words such as 'some' and 'lot' are top indicators for general sentences. Words indicative of specific sentences are 'a', 'one' and pronouns. However, a large number of other words appear to be domain specific indicators—'officials', 'number', 'prices' and 'business' for general sentences, and 'co.', 'inc' for the specific category.

The weights associated with non-lexical features conformed to our intuitions. Mentions of numbers and names are predictive of specific sentences. Plural nouns are a property of general sentences. However, the dollar sign, which we expected is more likely with specific sentences turned out to be more frequent in the other category. As for the language model features, general sentences tended to have lower probability and higher perplexity than specific ones. General sentences also have greater counts of polarity words (normalized by length) and higher number of adjectives and adverbs and their phrases. At the same time, these sentences have fewer and shorter verb phrases and fewer prepositional phrases.

3 Testing the classifier on new sentences

So far, we have used discourse relations as sources of general and specific sentences. Here we present

an annotation study where we asked people to directly judge a sentence as general or specific. We use these annotations to validate our classifier based on discourse relations and to ascertain whether these distinctions can be performed intuitively by people. So we only elicit annotations on a small set of examples rather than build a large corpus for training.

We also use sentences from articles from different news sources, enabling us to study the robustness of the classifier for news in general, beyond the more domain specific materials from the Wall Street Journal. Further we highlight a useful aspect of our predictions. We find that the confidence (probability from logistic regression) with which our classifier predicts the class for a sentence is correlated with the level of annotator agreement on the sentence. This finding suggests that the confidence scores can be used successfully to assign a graded level of specificity.

3.1 Annotations for general/specific

For our initial study outlined above, we have used the Instantiation and Specification sentences from Wall Street Journal texts in the PDTB. So we chose three WSJ articles from the PDTB corpus for further annotation, each around 100 sentences long. These articles were the ones with maximum number of Instantiations because we wanted to test whether people would judge the two sentences in Instantiations in the same manner as we have used them (the first general and the second specific).

We also chose articles from another corpus, AQUAINT (Graff, 2002), to compare the effect of corpus specifics on the classifier performance. These are a set of 8 news articles, six published by the Associated Press (AP) and two by Financial Times and are around 30 sentences each. Overall, there were 294 sentences from the WSJ and 292 from AP. Both sets of articles are about news, but the WSJ contains mainly financial reports.

We used Amazon's Mechanical Turk (MTURK)[3] to obtain annotations. We presented a user with one sentence at random and three options for classifying it: general/ specific/ can't decide. We provided minimal instructions[4] and annotators were encouraged to use their intuition to choose a judgement.

	WSJ articles			AP articles		
Agree	**total**	**gen**	**spec**	**total**	**gen**	**spec**
5	96	51	45	108	33	75
4	102	57	45	91	35	56
3	95	52	43	88	49	39
undecided	1			5		
Total	294	160	133	292	117	170

Table 4: Annotator agreement

We obtained judgements from 5 unique users for each sentence. However, it is not the case that all sentences were judged by the *same* 5 annotators. So we do not compute the standardly reported Kappa measures for annotator agreement. Rather, we present statistics on the number of sentences split by how many annotators agreed on the sentence class. We also indicate the number of sentences where the majority decision was general or specific (Table 4).

As we can see from the table, there were only very few cases (6 out of $\sim$600) where no majority decision was reached by the 5 annotators. For about two-thirds of the examples ($\sim$400) in both the WSJ and AP, there was either full agreement among the five annotators or one disagreement. These results are high for a new task where annotators mainly relied on intuition. Some examples of sentences with different agreement levels are shown in Table 5.

Here we can notice why the examples with low agreement could be confusing. Sentence [S2] judged as specific (by three annotators) contains details such as the exact quantities of rainfall. At the same time, it contains the vague phrase "still had only". The remaining two annotators could have seen these general properties as more relevant for their judgement. Sentence [G2] which also had low agreement, has some general properties but also specific information, such as the phrase describing the word "companies".

In terms of the distribution of general and specific sentences, the two sets of articles differ. In the WSJ, there are more general (55% of total) than specific sentences. In the AP articles, specific sentences form the majority (60%) and there is a wider gap between the two types. One reason for this difference could be the length of the ar-

[3] http://sites.google.com/site/amtworkshop2010/

[4] "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 made 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."

General	Agree = 5	[G1] The conditions necessary for a dollar crisis had been building up in currency markets for some time.
	Agree = 3	[G2] The flip side of the hurricane's coin was a strong showing from the stocks of home construction companies expected to benefit from demand for rebuilding damaged or destroyed homes.
Specific	Agree = 5	[S1] By midnight, 119 mph winds were reported in Charleston.
	Agree = 3	[S2] But the weather service said all Mississippi farm communities still had only 30 percent to 50 percent of normal moisture.

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

	General	**Specific**
Sent1	29 L5(14), L4(9), L3(6)	3 L5(1), L4(1), L3(1)
Sent2	6 L5(1), L4(3), L3(2)	26 L5(13), L4(9), L3(4)

Table 6: Annotator judgements on instantiation sentences

	General	**Specific**
Sent1	10 L5(4), L4(3), L3(3)	6 L5(1), L4(1), L3(4)
Sent2	8 L5(5), L4(3), L3(0)	8 L5(5), L4(2), L3(1)

Table 7: Annotator judgements on specification sentences

ticles. Those from WSJ are much longer than the AP articles and probably longer articles have more topics and corresponding general statements.

3.2 Results on Instantiation examples

We have assumed from the definitions of Instantiation and Specification relations, that their first sentences (*Sent1*) are general and their second (*Sent2*) specific. Further, we used these two sentences independently in two different classes. Now we test this intuition directly. Would people given only one of these sentences in isolation, give it the same judgement of generality as we have assumed?

There were 32 Instantiations and 16 Specification relations in the three WSJ articles we annotated and each of these relations is associated with two sentences, *Sent1* and *Sent2*. In Tables 6 and 7, we provide the annotator judgements and agreement levels on these sentences. The number of sentences x in each category with a certain level of agreement y is indicated as Ly(x). So L5(3) means that three sentences had full agreement 5.

For Instantiations, we find that the majority of *Sent1* are judged as general and the majority of *Sent2* are specific, 80% in each case. But for both *Sent1* and *Sent2*, there is one sentence which all the annotators agreed should be in the opposite class than assumed. So there are some cases where without context, the judgement can be rather different. But such examples are infrequent in the Instantiation sentences.

On the other hand, Specifications show a weaker pattern. For *Sent1*, still a majority (62.5%) of the sentences are called as general. However, for *Sent2*, the examples are equally split between general and specific categories. Hence it is not surprising that the Instantiation sentences have more detectable properties associated with the first general sentence and the second specific sentence and the classifier trained with these examples obtains better performance compared with training on Specifications.

3.3 Classifier accuracy and confidence

Now we test our classifier trained on Instantiation relations on the new annotations we have obtained on WSJ and AP articles. The parse trees for sentences in the test set were obtained using the Stanford Parser (Klein and Manning, 2003). Since our classifier was trained on Instantiations sentences from the WSJ, when testing on the new WSJ annotations, we retrained the classifier after excluding sentences that overlapped with the test set.

Our goal here is to a) understand the performance and genre independence of our features on the new test set b) explore the accuracy on examples with different levels of annotator agreement c) build a combined classifier using both discourse relations and direct annotations.

In each line of Table 8, we report the performance on examples from the specified agreement levels. A '+' sign indicates that examples from multiple agreement levels were combined.

Non-lexical features give the best performance on both sets of articles. The word features trained on WSJ Instantiations give more than 10% lower accuracy than non-lexical features, even on the WSJ articles. So lexical features probably do not cover all example types but non-lexical features provide better abstraction and portability across corpora. The accuracy of the non-lexical features on all directly annotated examples (*Agreement 3+4+5*) is 76% on WSJ and 81% on AP, similar to results on the Instantiation sentences.

But the accuracy increases on examples with

Examples	WSJ sentences Size	All features	Nonlexical	Words	AP sentences Size	All features	Nonlexical	Words
Agreement 5	96	90.6	96.8	84.3	108	69.4	94.4	78.7
Agreement $4+5$	198	80.8	88.8	77.7	199	65.8	89.9	74.8
Agreement $3+4+5$	293	73.7	76.7	71.6	287	59.2	81.1	67.5

Table 8: Accuracy of classifier on annotated examples

higher agreement and is over 90% for sentences with full agreement. The sentences with more agreement appear to have easily detectable properties for the respective class and so the classifier produces accurate predictions for them. As we saw, examples with low annotator agreement (Table 5) probably have a mix of properties from both classes. We further analyze the relationship between agreement and classifier performance by studying the classifier confidence scores.

In Table 9, we report the mean value of the classifier confidence for predicting the *correct class* for sentences having different agreement levels. A correct prediction occurs when the confidence is above 0.5 for the target class, so all the values we consider here are above 0.5. We now want to study when the correct prediction is made, how large is the confidence on examples with different annotator agreement levels. When the mean value of confidence scores at a particular agreement level was significantly better than another (determined by a two-sided t-test), those levels with lower confidence are indicated within parentheses.

As expected, the confidence of the classifier is significantly higher at greater levels of agreement again proving that the examples with higher annotator agreement are easier to classify automatically. So, the probability value produced by the classifier could be a better metric to use than the hard classification into classes. Further, since humans do have a low agreement on one-third of the sentences, a graded value is probably more suitable for the prediction of generality of a sentence.

We now have a larger set of annotated examples, so we combine the sentences from these two corpora with the Instantiation examples and build a combined classifier. Here the total general sentences is 1648 and there are 1674 specific sentences. So the distribution is almost equal and the baseline random performance would be 50% accuracy. The 10-fold cross validation accuracies from non-lexical, word and 'all features' on this full set are shown below.

Nonlexical : 72.36
Words : 72.36
All features: 74.68

Agreement	WSJ	AP
5	0.77 (4, 3)	0.78 (4,3)
4	0.70	0.70 (3)
3	0.67	0.66

Table 9: Mean value of confidence score on correct predictions

Here, after combining the examples, the classifier learns the lexical features indicative of both types of articles. So we end up with a similar trend as on the Instantiations based classifer. Both non-lexical and word features individually obtain 72% accuracy. Their combination is slightly better with 75% accuracy. So word features only when trained on both types of data again end up becoming good predictors and complementary with non-lexical features. So for new domains, the non-lexical features would be more robust.

Overall, we have provided a classifier that has high accuracy on a diverse set of examples.

4 Task based evaluation

So far we have tested our classifier on individual sentences which were judged as general or specific. Now we provide a task-based evaluation on news summaries. Here people were asked to write general or specific summaries for a set of articles, in the first type conveying only the general ideas and in the second providing specific details about the topic. We show that our classifier successfully distinguishes these two types of summaries.

Summarization is one task where the distinction between general and specific content is relevant. The space available for summary content is limited. So authors include some specific detail but at the same time have to generalize other content to stay within the space limit. Early work in Jing and McKeown (2000) report that when people create summaries, they generalize some of the sentences from the source text, others are made more specific. From the point of view of automatic systems, Haghighi and Vanderwende (2009) developed a topic model-based summarization system which learns the topics of the input at both overall document level as well as specific subtopics. Sentences are assumed to be generated by a combination of the general and specific topics in the input

texts. However, since the preference of such sentences is not known, only heuristics were applied to choose the proportions. We expect our classifier to be useful in such cases.

4.1 Data

We use summaries and source texts from the Document Understanding Conference (DUC) organized by NIST in 2005.[5] The task in 2005 was to create summaries that are either general or specific. Each input consists of 25 to 50 news articles on a common topic. A topic statement is provided for *each input* which states the user's information need. Gold standard summaries for evaluation are created by human assessors for all these inputs. A length limit of 250 words is enforced.

During the creation of input sets, the annotators were asked to specify for each input, the type of summary that would be appropriate. So annotators provided a desired *summary granularity* for each input: either *general* or *specific*. There were a total of 50 inputs, 24 of them were marked for general summaries, the remaining for specific.

Next these input texts and topic statements were given to trained NIST assessors for writing summaries.[6] For some inputs (20), 9 summaries each were provided by the assessors, other inputs had 4 summaries. Considering the granularity of inputs, there is a roughly equal distribution of general (146) and specific (154) summaries. We now test if our classifier predictions can distinguish between these general and specific summaries where people relied on an intuitive idea of general and specific content overall in the summary.

4.2 Difference in specificity

For this analysis, we use the combined classifier from the Instantiation relations and extra annotations. We used the combination of all features since it gave the best performance for this setup.

Next we assigned a specificity level for each summary in the following way. For each sentence in the summary, we obtained the classifier confidence for predicting the sentence to be "specific". Each token in the summary was assigned the confidence of the sentence in which it appeared. Then the *average specificity of words* in the summary was computed as the average value of this confidence measure over all the tokens in the summary.

[5] http://duc.nist.gov/duc2005/

[6] The guidelines and example summaries can be found at http://duc.nist.gov/duc2005/.

Text	General category	Specific category
Summaries	0.55 (0.15)	0.63 (0.14)
Inputs	0.63 (0.06)	0.65 (0.04)

Table 10: Mean value (and standard deviation) of specificity levels for inputs and summaries

The statistics for this score in the general and specific categories are shown in Table 10.

For specific summaries, the mean specificity is 0.63, while for general ones it is only 0.55. The difference is also statistically significant under a two sided t-test (p-value of 1.5e-06). This result shows that our predictions are able to distinguish the two types of summaries.

We also computed the specificity scores for inputs in the same manner. Here the mean value is around 0.63 and does not vary significantly between the two classes (pvalue = 0.275). So while the inputs do not vary in specificity for the two categories, the summary authors have injected the required granularity during summary creation. To emulate the human summaries, systems would need to optimize for a measure of general/specific rather than use a generic strategy. Our classifier's predictions could be combined with other content selection features for such purposes.

5 Conclusion

We have introduced a new task—identification of general and specific sentences. We have shown how certain discourse relations involve these two types of sentences and can be used as training data for the task. We introduced features such as polarity, word specificity, language models, entity-related and lexical features which resulted in high classification performance, 25% absolute increase over the baseline. Our classifier also provides a graded score for specificity and can distinguish general and specific summaries written by people.

With this success, for future work, we plan to investigate the use of our classifier in applications which can use the general/specific notion. One task is providing feedback during writing. By learning patterns of use of general and specific sentences, we can use our predictions to annotate sentences which need more support from the writer. We also plan to explore pairs of general and specific sentences for the task of question generation. Specific sentences with important content can be treated as a potential answer, while a general sentence on the same subtopic can be used to generate the question.

References

G.J. Alred, C.T. Brusaw, and W.E. Oliu. 2003. *Handbook of technical writing*. St. Martin's Press, New York.

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

A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of NAACL-HLT*, pages 362–370.

H. Jing and K. McKeown. 2000. Cut and paste based text summarization. In *Proceedings of NAACL*.

H. Joho and M. Sanderson. 2007. Document frequency and term specificity. In *Proceedings of RIAO*.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430.

D. Marcu and A. Echihabi. 2001. An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL*, pages 368–375.

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

T. Mathew and G. Katz. 2009. Supervised categorization for habitual versus episodic sentences. In *Sixth Midwest Computational Linguistics Colloquium. Indiana University Bloomington, May*, pages 2–3.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to wordnet: An online lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312.

R. Prasad, E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, L. Robaldo, and B. Webber. 2007. The penn discourse treebank 2.0 annotation manual. *http://www.seas.upenn.edu/ pdtb*.

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*.

N. Reiter and A. Frank. 2010. Identifying generic noun phrases. In *Proceedings of ACL*, pages 40–49.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pages 448–453.

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

C. Sporleder and A. Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14:369–416.

P.J. Stone, J. Kirsh, and Cambridge Computer Associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347–354.