Improving the Annotation of Sentence Specificity

Junyi Jessy Li[1], **Bridget O'Daniel**[2], **Yi Wu**[1], **Wenli Zhao**[1], **Ani Nenkova**[1]
[1] University of Pennsylvania, Philadelphia, PA 19104
{ljunyi,wuyd,wenliz,nenkova}@seas.upenn.edu
[2] Berea College, Berea, KY 40404
Bridget_Odaniel@berea.edu

Abstract

We introduce improved guidelines for annotation of sentence specificity, addressing the issues encountered in prior work. Our annotation provides judgements of sentences in context. Rather than binary judgements, we introduce a specificity scale which accommodates nuanced judgements. Our augmented annotation procedure also allows us to define where in the discourse context the lack of specificity can be resolved. In addition, the cause of the underspecification is annotated in the form of free text questions. We present results from a pilot annotation with this new scheme and demonstrate good inter-annotator agreement. We found that the lack of specificity distributes evenly among immediate prior context, long distance prior context and no prior context. We find that missing details that are not resolved in the the prior context are more likely to trigger questions about the reason behind events, "why" and "how". Our data is accessible at `http://www.cis.upenn.edu/%7Enlp/corpora/lrec16spec.html`

Keywords: specificity rating, underspecification, discourse

1. Introduction

Louis and Nenkova (2012) introduced a corpus of sentences annotated as general or specific. Their definition of sentence specificity relied mostly on examples and intuition, related to the amount of detail contained by the sentence. They used the corpus of general and specific sentences to evaluate a classifier for the binary task (Louis and Nenkova, 2011a) and showed that changes in sentence and overall text specificity are strongly associated with perceptions of text quality. Science writing of the best quality in the New York Times is overall more general than regular science pieces in NYT and contain fewer stretches of specific content (Louis and Nenkova, 2013). Automatic summaries, which are often judged to be incoherent, are significantly more specific than same length human-written summaries for the same events (Louis and Nenkova, 2011b). Sentence specificity is also more robust than sentence length as indicator of which sentences may pose comprehension problems and need to be simplified for given audiences (Li and Nenkova, 2015). It is also a stable predictor in identifying high-quality arguments in online discussions (Swanson et al., 2015).

Given the demonstrated importance of sentence and text specificity in practical applications and the known shortcomings of the existing annotation, we set out to develop a more detailed framework for annotation of sentence specificity. In the brief annotation guidelines of Louis and Nenkova (2012), the general vs. specific distinction was defined in the following way:

> "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."

Their analysis of annotator disagreement led to the conclusion that a scale of specificity would be more appropriate than trying to classify sentences in two strict classes (general/specific) and that context information should be incorporated in the annotation to resolve anaphoric and topical references that otherwise appear insufficiently specific. In this work, we present a pilot corpus for contextually informed sentence specificity that enables the joint analysis of the *degree*, *location* and *manner* of underspecification in text:

- **Degree**: the specificity of a sentence is judged on a scale rather than as a binary factor as in Louis and Nenkova (2012);
- **Location**: segments that lack specificity are marked within each sentence;
- **Manner**: the cause of underspecification is provided for each marked segment, along with their relationship with prior context.

2. A more specific definition of sentence specificity

The aim in developing the new annotation scheme was to make more explicit what it means for a sentence to "stand on its own", while still keeping it general enough to solicit judgements from lay annotators. A sentence stands on its own if the semantic interpretation of referents can be easily disambiguated by a reader to that of the intended referent, the truth value of statements in the sentence can be determined solely based on the information in the sentence and commonly shared background knowledge, and key information about the participants and causes of an event are fully expressed in the sentence.

These three requirements cover a broad range of linguistic and semantic phenomena. For example a reference to a discourse entity may not be readily interpretable when the reference is anaphoric, by either a pronoun or definite noun phrase, when the reference is by proper name

with which the reader is not familiar or the reference is generic, not referring to a specific discourse entity at all (Dahl, 1975; Reiter and Frank, 2010). Similarly gradable adjectives (Frazier et al., 2008; de Marneffe et al., 2010) like "tall", "smart" and "valuable" are interpreted according to an assumed standard. If the standard is unknown or if the writer and the reader do not share the same standard for interpreting these properties, it is impossible to verify if a sentence has the same truth value for both the writer and reader. These issues of ability to verify the truth value of a statement are directly related to Wiebe (2000)'s original definition of adjective subjectivity. Sentences like "He is a publishing sensation" and "He is a valuable member of our team" are subjective because different people's definitions of what selling records are sensational or what constitutes a valuable member may differ radically. Similarly when a typical argument of a verb is missing from a sentence (Palmer et al., 2005), the reader may have difficulty understanding the full event that is being described.

Word choice can also determine the overall specificity of a sentence, by making more explicit the manner in which an action is performed or the identity of the discourse entity, as shown by the contrast of sentence pairs like "The worker cleaned the floor" vs. "The maid swept the floor" (Stinson and Tracy, 1983; Resnik, 1995; McKinlay and Markert, 2011; Nastase et al., 2012).

The annotation we propose indirectly provides mechanisms to analyze which of the above intricate linguistic and semantic phenomena trigger the need for clarification of naive readers interested in gaining good understanding of a text. It is developed with the flexibility and intention to enable further analysis such as the classification of triggers and future refinement of annotation, to provide a practical connection between language-related applications and linguistic phenomena.

3. Methodology and corpus summary

The annotation is carried out on news articles. Each article is divided into groups of 10 consecutive sentences that the annotators would work on in one session. If the selected text was found in the middle of an article, the previous sections of the article were provided to the annotators at the start of the task for reading, but participants were not asked to annotate them.

For each sentence, the annotators rate its specificity based on a scale from 0 - 6 (0 = most specific: does not require any additional information to understand who or what is involved and what is the described event; 6 = most general). For this judgement, annotators consider each sentence independent of context.

Then they mark text segments that are underspecified, identify the cause of underspecification in the form of free text questions, and identify if these questions may be answered by information given in previous context. If the annotator chose not to ask any question, she is asked to distinguish if the sentence is most specific (i.e., no underspecified segments) or most general (i.e., the sentence conveys general information that needs no further specification). The latter types of sentences capture generics such as "Cats have four paws." that do not refer to specific events or entities (Carlson, 2005). Agreement on annotating generic noun phrases is low (Nedoluzhko, 2013), so we adopt a more high-level annotation at the sentence level that can be done with less training and with higher agreement.

There are four types of status concerning previous context:

- **In the immediate context**: the answer to the question can be found in the two immediately preceding sentences, a distance shown to be the median length of pronoun chains in writing (Hindle, 1983). Here we use this as the effective context for pronoun resolution.
- **In some previous context**: the answer to the question can be found in the article but it is in a sentence more than two sentences before the one currently being annotated.
- **Topical**: the answer is not explicitly given in the preceding discourse but can be inferred from it.
- **None**: the answer is not explicitly or implicitly included in the preceding discourse. The author intentionally left it unspecified or it is specified in the following discourse.

Additionally, we ask the annotators to only ask questions that need to be answered in order for them to properly understand the sentence and to mark only the minimal span in the sentence which needs further specification. For example,

> **[sentence]** He sued the executive of the company.
>
> **[question]** "sued": Why did he sue? (Topical).

The annotator chose the word "sued" rather than "He sued" or "He sued the executive" because the question only relates to the act of suing.

The annotators are native speakers of North American English (one Canadian and two Americans). The annotation is performed on 16 articles from the New York Times dataset (Sandhaus, 2008) (13 out of the 16 are full article annotations; the annotations are all carried out from the beginning). Eight of these are politics articles and the other eight business articles. A total of 543 sentences and 15,224 words were triple annotated by each of the annotators. The annotators generated 2,796 questions.

4. Specificity ratings

We first present the overall distribution of sentence specificity, then go into analysis of annotator agreement of specificity ratings.

Sentence specificity distribution. We compute the sentence specificity score as the average from the ratings from all three annotators. Higher scores indicate more general sentences. As shown in Figure 1, the distribution of the ratings is roughly normal, with mean at the slightly general side. In other words most sentences are a mixture of general and specific information, confirming the need for a rating scheme rather than a binary one.

Agreement. We first compute the standard deviation of ratings for the sentences in the corpus. Notably, 90.4% of

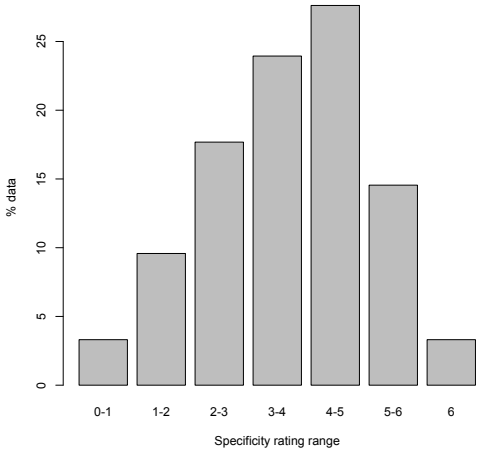

Figure 1: Sentence specificity distribution.

	human	random
Cronbach α	0.7224	0.4886

Table 1: Cronbach's alpha for sentence specificity, random vs. human.

the standard deviation is below 1 and 64.3% below 0.5, indicating that the ratings for each sentence are close to one another.

In Table 1, we show Cronbach's α (Cronbach, 1951) for annotators' consistency and that for informed random ratings. To generate the random ratings, we randomly draw a rating from the multinomial distribution given by the overall sentence specificity distribution shown in Figure 1 for each sentence. This process is repeated 1,000 times and the αs are averaged. Human ratings are much higher than the randomly generated ones.

Values of Cronbach's α are usually interpreted as good when its values are larger than 0.8, acceptable when its values are in the 0.7–0.8 range and unacceptable when lower than 0.5. According to this interpretation, human ratings exhibit acceptable agreement while the randomly generated ones are unacceptable.

Appendix A and B show examples of sentences with high and low agreement on specificity ratings.

We also compute specificity rating agreement on the document level. The specificity of a document is computed as the average of the specificity ratings of the sentences in it. The correlation of document specificity scores is very high, equal to 0.98 for all three pairs of annotators.

5. Question location and type

In this section, we analyze annotator agreement on identifying the location of a sentence segment that requires further specification for complete understanding of the sentence. We also tabulate the type of questions that were asked regarding the missing information. The annotators are asked to mark out the minimal text span for which she needs further specification. Each segment is associated with a free-text question and the location of the answer is given as one of *immediate context*, *previous context*, *topical*, or *none*.

non-overlap	overlap	containment
0.3	29.8	69.9

Table 2: % of questions for the three states.

1	2	3
60.0	26.2	13.8

Table 3: % of underspecified tokens marked by 1, 2 or all 3 annotators.

5.1. Consensus on underspecified segments

The annotators asked 2,796 questions, each associated with a sentence substring (span) which the annotator identified as needing further specification. We consider three possible states for sentence substrings marked by different annotators: containment, overlap and non-overlap. Let the span of question q_i be s_i. For each question, we first check for containment among all other questions in the same sentence: $\forall j,\ s_i \in substring(s_j) \vee s_j \in substring(s_i)$. If not, we look for an overlap: $\forall j,\ s_i \cap s_j \neq \emptyset$. If neither containment nor overlap is found, we assign the "non-overlap" state to the question.

Table 2 shows the percentage of questions with each state. It confirms that when an annotator identifies an underspecified segment, it is 99.7% likely that a part or all of the segment is also identified as underspecified by at least one other annotator. This means that the readers reach a natural consensus as of which part of the sentence needs further detail. Furthermore, the majority (69.6%) of these segments fully overlap with another.

We also calculate the % of underspecified tokens that are marked by one, two or all three annotators, tabulated in Table 3. Despite the high overlap of *segments* demonstrated above, there is a high percentage of tokens marked by only one annotator. This shows that despite the minimality principle, identifying underspecified *tokens* of high agreement requires additional filtering.

5.2. Sub-sentential vs. sentential specificity

Since the annotators are asked to give specificity ratings and ask questions independently, we can now compare number of underspecified segments at the sub-sentence level with the specificity of the overall sentence. For the former, we calculate in each sentence, the % of tokens marked as underspecified by at least one annotator. If an annotator did not ask a question and marked the reason to be that the sentence is too general, then a count 1 is added to all tokens in the sentence. Figure 2 shows that the more general the sentence was judged to be, the larger its portion of underspecified tokens.

5.3. Discourse analysis of underspecification

To support understanding of document level information packaging, we link each sub-sentential underspecified text segment with the running discourse by annotating the location of answers to the question associated with each segment. Here we show the % of questions whose answers can be found in the immediate context, in previous context, is

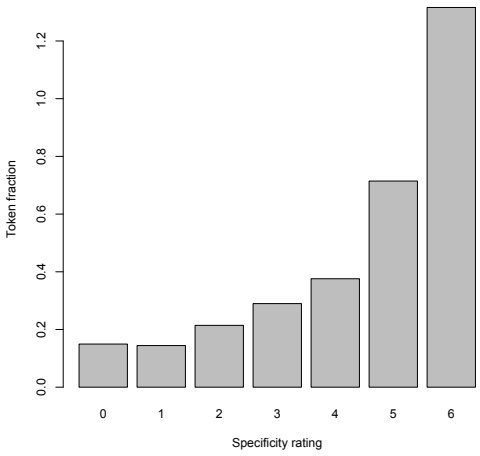

Figure 2: Average fraction of tokens marked as underspecified vs. overall sentence specificity.

Immediate	Previous	Topical	None
32.47	30.87	7.37	28.97

Table 4: % of questions and its context status.

topically related, or not in any prior context. Table 4 shows that the context status of underspecification is remarkably even in the none, immediate and previous context cases, with a small portion being topical.

The type of question—signaled by the question word—gives insight to what type of information a reader is seeking to understand the sentence. The context status of the question gives information for each segment where it can be specified in the running discourse. In Table 5, we tabulate the number of interrogatives found in the questions along with the context status associated with each interrogative, sorted by the frequency of the interrogative. The most frequent interrogative is "what", followed by "who", "how", "why" and "which"; "where" and "when" questions are not often raised by the annotators[1]. These question words also distribute very differently in each context status; for example, most of the underspecification leading to "what", "who", "which" and "where" questions can be resolved in prior context, but "how", "why" and "when" questions are raised mostly when the lack of specificity cannot be resolved in prior context.

6. Linguistic analysis

In this section, we show three aspects of analysis that highlights the wider implications of our sentence specificity annotation to broader questions in computational linguistics and semantics.

[1]Note that interrogatives and question types do not have a one-to-one mapping. For example, not all "what" questions are entity-centric. We found 186 of these questions that are potentially causal questions, with presence of the words *happen, reason, for, cause, mean, entail, purpose*. We leave for future work a detailed classification of question types.

Interrogative	All	Immediate	Previous	None
what	1388	36.6	36.5	20.0
who	419	52.7	31.7	11.5
how	332	4.5	10.2	76.2
why	317	10.4	24.3	50.5
which	242	40.9	35.2	21.1
where	66	36.4	37.9	22.7
when	24	20.8	12.5	62.5

Table 5: Number of question interrogatives and percentages of associated context status.

POS tag	Specified	Immediate	Previous	None
ADJ	69.0	5.7	6.3	19.8
ADP	93.7	1.8	1.6	2.8
ADV	72.6	7.9	5.2	14.9
CONJ	93.1	2.1	1.9	3.7
DET	75.8	9.1	5.6	9.8
the	68.5	12.0	14.4	6.8
NOUN	71.32	7.8	12.7	10.3
NUM	88.29	5.1	4.8	3.2
PRON	67.3	21.8	12.2	1.9
PRT	90.4	2.1	2.1	4.5
VERB	82.5	3.6	4.6	9.4

Table 6: Percentage of part of speech usage in specified and underspecified segments, stratified by associated context status.

6.1. Specificity and content density

The annotation of articles using a scale of specificity score allows us to study the connection between text specificity and content density. The latter, described in Yang and Nenkova (2014), represents how much the text is factual and how well the content is expressed in a "direct, succinct manner".

Specifically, our articles overlap with those annotated in Yang and Nenkova (2014), so we compare the content density scores of lead paragraphs annotated by Yang and Nenkova (2014) with their specificity. For each lead paragraph, its content density is a real-valued score assigned by two annotators (here we take the average). A larger value indicates more density. Its specificity score is obtained by averaging the sentence specificity ratings (for each sentence its specificity rating is averaged among annotators). We observe a significant ($p \leq 0.05$) Spearman correlation of -0.51, indicating that content-density on the paragraph level is positively associated with its sentences being more specific.

6.2. Underspecified tokens and context

Previously we observed that about a third of the lack of specificity cannot be resolved in prior context. Here we offer insight into the characteristics of the tokens associated with this category of underspecification. In Table 6, we lay out the percentage of universal part-of-speech tags (Petrov et al., 2012) of tokens in underspecified segments their percentage associated with the following: fully specified, resolved in immediate context, in previous context and no context. We also separated the definite determiner "the" from the main determiner category to distinguish between definite and indefinite references. Each token is counted

once if marked by multiple annotators.

These numbers clearly show that most of the underspecification comes from content words. Among them, most of the lack of specificity of pronouns and determiners can be resolved in prior context. The definite expression "the" behaves differently from non-definites; it is one of the most often marked POS tags (and most of them can be resolved in context), while other determiners are marked much less often, with a large portion that cannot be resolved in context. On the other hand, the lack of specificity from adjectives, adverbs and verbs more often cannot be resolved in context.

This information when combined with interrogative breakdown in Table 5 illustrates that underspecified content is more likely to be non-entities and triggers high level comprehension questions.

6.3. Entity co-reference

We analyzed the connection between co-reference resolution and context-dependent underspecification (questions about content missing in the sentence but found in preceding context and necessary for full comprehension of the sentence). It is reasonable to assume that all questions resolved in the previous context involved anaphoric references to previously mentioned entities. Yet, of the underspecified segments annotated as having the missing details in the local context (i.e., two sentences above), only 34.4% contain an entity that is resolved by automatic coreference resolution[2]. For non-local previous context, this number is 26% (21.5% for all segments). This confirms that our corpus captures information packaging patterns beyond noun phrase anaphora resolution problems. An illustrative example is shown below:

> After a contest that had pitted domestic pride against global politics, the Pentagon yesterday chose an international team, headed by Lockheed Martin, to build the next fleet of presidential helicopters over Sikorsky Aircraft, which had positioned itself as the "all-American" choice. *In selecting Lockheed, which will receive $ 1.7 billion initially to begin the program, the Pentagon signaled a new openness to foreign partners on sensitive military tasks.*
> **Question:** "selecting" — What were they selected for? (immediate context)

7. Conclusion

In this work, we present an annotation method and a corpus for context-informed sentence specificity. Our methodology enables joint annotation on sentential specificity, subsentential underspecified expressions and their context dependency. We annotate the type of underspecification using high level questions generated by the annotators. We showed that the annotators reached good agreement on sentence and document level specificity and they have high consensus as which text segments within the sentence are underspecified. We plan to release our dataset and further expand it to enable more sophisticated linguistic analysis.

[2]We used the Berkeley Entity Resolution System (Durrett and Klein, 2014).

Appendix: examples

Formatting. Each example includes the sentence to be rated in *italic* as well as the two consecutive sentences immediately before (i.e., immediate context). The ratings are shown in

> annotator:rating

format. For questions, they are formatted as:

> "underspecified text" — question body (context status)

A High agreement

[Ex1: general] Those forces are made up of about 150,000 troops from the United States and upward of 25,000 from other nations. But Dr. Allawi raised the tantalizing prospect of an eventual American withdrawal while giving little away, insisting that a pullout could not be tied to a fixed timetable, but rather to the Iraqi forces' progress toward standing on their own. *That formula is similar to what President Bush and other senior administration officials have spoken about.*
Ratings: A1:5, A2:5, A3:5
Questions:
Q1: "That formula" — What is the formula? (immediate context)
Q2: "similar" — How similar? (no context)

[Ex3: specific] The deaths of two Americans announced by the United States military on Friday — a marine killed by gunfire in Falluja and a soldier killed by a roadside bomb in Baghdad — brought the total killed since the war in Iraq began in March 2003 to 2,178. The total wounded since the war began is 15,955. *From Jan. 1, 2005 to Dec. 3, 2005, the most recent date for which numbers are available, the number of Americans military personnel wounded in Iraq was 5,557.*
Ratings: A1:0, A2:0, A3:0
Questions:
Q1: "in Iraq" — What was the conflict? (no context)

[Ex4: underspecification (in context)] The aircraft was the first large plane acquired by the new Iraqi Air Force, which was one of the most powerful in the Middle East before it was decimated by bombing attacks in the 1991 Persian Gulf war. Dr. Allawi's remarks were made on a day when attacks underlined, once again, how insurgents have turned wide areas of the country, including Baghdad, into what is effectively enemy territory, with an ability to strike almost at will, and to shake off the losses inflicted by American troops. *The attacks in Baghdad on Wednesday were aimed at the approaches to the Australian Embassy and four Iraqi security targets, including a police station, an army garrison and a bank where policemen were lining up to receive their monthly pay.*
Ratings: A1:1, A2:1, A3:1

Questions:
Q1: "attacks" —How were they attacked? (previous context)
Q2: "The attacks" — What were the attacks? (previous context)
Q3: "The attacks" — Who was responsible? (previous context)

[Ex5: underspecification (no context)] Sales increased 4.5 percent, to 30.3 billion euros. The results for Philips and its plan to buy back shares drove its share price up 4.45 percent, to 19.50 euros. *Jan Hommen, the chief financial officer, said that Philips believed its stock was undervalued.*
Ratings: A1:4, A2:4, A3:4
Questions:
Q1: "undervalued" — Why? (no context)
Q1: "undervalued" — By how much? (no context)
Q1: "undervalued" — Why did they believe it was undervalued? (no context)

B Low agreement

[Ex1] Right now, there are around 425 attacks a week in Iraq, according to American officials, nearly the same rate as six months ago. The newly released archbishop, Basile Georges Casmoussa, was in a jovial mood as he met with well-wishers at his Eastern Rite Catholic church in Mosul, a crime-ridden city in northern Iraq that contains one of Iraq's larger Christian communities. *Christians make up about 3 percent of Iraq's total population.*
Ratings: A1:2, A2:0, A3:6
Questions:
Q1: "Christians" — What kind of Christians? (no context)
Q2: "total population" — What is the total population? (no context)

[Ex2] A year ago, about 25 percent of attacks inflicted casualties. More than 400 car and suicide bombs struck the country in 2005, although the number has dropped sharply in recent months. *In April, for instance, there were 66 suicide and car bomb attacks, compared with 28 in November.*
Ratings: A1:4, A2:0, A3:1
Questions:
Q1: "for instance" — Instance of what? (immediate context)
Q2: "car bomb attacks" — Who is perpetrating these attacks? (previous context)

[Ex3] The second worst month was October, when 96 Americans were killed and 603 wounded. More than half of all 2005 American military deaths, 427, were caused by homemade bombs, most planted along roadsides and detonated as vehicles passed. *American commanders have said that roadside bombs, the leading cause of death in Iraq, have grown larger and more sophisticated.*
Ratings: A1:1, A2:5, A3:4
Questions:
Q1: "more sophisticated" — How are they more sophisticated? (no context)
Q2: "have grown larger and more sophisticated" — How so? (no context)

Ackowledgements

We thank Muffy Siegel and the reviewers for their valuable comments and feedback.

References

Carlson, G. (2005). Generic reference. *In The Encyclopedia of Language and Linguistics, 2nd Ed. Elsevier*.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Dahl, O. (1975). On generics. *Formal Semantics of Natural Language*, pages 99–111.

de Marneffe, M.-C., Manning, C. D., and Potts, C. (2010). "Was it good? It was provocative." Learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 167–176.

Durrett, G. and Klein, D. (2014). A joint model for entity analysis: Coreference, typing, and linking. In *Transactions of the Association for Computational Linguistics (TACL)*.

Frazier, L., Jr., C. C., and Stolterfoht, B. (2008). Scale structure: Processing minimum standard and maximum standard scalar adjectives. *Cognition*, 106(1):299 – 324.

Hindle, D. (1983). Discourse organization in speech and writing. In Muffy E. A. Siegel et al., editors, *Writing Talks*. Boynton/Cook.

Li, J. J. and Nenkova, A. (2015). Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, pages 2281–2287, January.

Louis, A. and Nenkova, A. (2011a). Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 605–613.

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

Louis, A. and Nenkova, A. (2012). A corpus of general and specific sentences from news. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Louis, A. and Nenkova, A. (2013). A corpus of science journalism for analyzing writing quality. *Dialogue and Discourse*, 4(2):87–117.

McKinlay, A. and Markert, K. (2011). Modelling entity instantiations. In *Recent Advances in Natural Language Processing, (RANLP)*, pages 268–274.

Nastase, V., Judea, A., Markert, K., and Strube, M. (2012). Local and global context for supervised and unsupervised metonymy resolution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 183–193.

Nedoluzhko, A. (2013). Generic noun phrases and annotation of coreference and bridging relations in the prague dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 103–111.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Reiter, N. and Frank, A. (2010). Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 40–49.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1 (IJCAI)*, pages 448–453.

Sandhaus, E. (2008). *The New York Times Annotated Corpus LDC2008T19*. Linguistic Data Consortium.

Stinson, M. and Tracy, O. A. (1983). Specificity of word meaning and use of sentence context by hearing-impaired adults. *Journal of Communication Disorders*, 16(3):163 – 173.

Swanson, R., Ecker, B., and Walker, M. (2015). Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 217–226.

Wiebe, J. (2000). Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (AAAI)*, pages 735–740.

Yang, Y. and Nenkova, A. (2014). Detecting information-dense texts in multiple news domains. In *Proceedings of the Twenty-Eighth Conference on Artificial Intelligence (AAAI)*.