Phrase Generalization: a Corpus Study in Multi-Document Abstracts and Original News Alignments

Ariani Di-Felippo
Federal University of São Carlos
Language and Literature Dept.
Rod. Washington Luis, km 235 (SP-310)
São Carlos, SP 13565-905, Brazil
arianidf@gmail.com

Ani Nenkova
University of Pennsylvania
Computer and Information Science Dept.
3330 Walnut St.
Philadelphia, PA 19104, USA
nenkova@seas.upenn.edu

Abstract

Content can be expressed at different levels of specificity, varying the amount of detail presented to the reader. The need to transform specific content into more general form naturally arises in summarization, where people and machines need to convey the gist of a text within imposed space constraints. Completely removing sentences and phrases is one way to reduce the level of detail. The bulk of work on summarization content selection and compression deal with these tasks. In this paper, we present a corpus study on a more subtle and understudied phenomenon: noun phrase generalization. Based on multi-document news and abstract alignments at the phrase level, we arrive at a five category classification scheme and find that the most common category requires semantic interpretation and inference. The others rely on lexical substitution or deletion of details from the original expression. We provide a systematic analysis, elucidating the capabilities needed for automating the generation of more general or more specific references.

1 Introduction

Summarization involves a number of complex transformations to condense the gist of a text into a short summary (Nenkova and McKeown, 2011). One of these transformations is changing the amount of detail in the original news texts. Removing entire sentences is one of the fairly well-understood ways for changing the amount of detail. Which sentences to remove can be decided in a system's content selection module by a number of competitive approaches (Gillick and Favre, 2009; Lin and Bilmes, 2011; Kulesza and Taskar, 2011). Similarly, one can perform sentence compression, removing words or phrases from a sentence in the original text to form a summary sentence (Knight and Marcu, 2000; Riezler et al., 2003; Turner and Charniak, 2005; McDonald, 2006; Galley and McKeown, 2007; Cohn and Lapata, 2008) or perform sentence selection and compression jointly (Berg-Kirkpatrick et al., 2011).

In this paper, we focus our attention on a much finer level to study the changes of specificity on the phrase level. The existence of these changes have been documented in prior work (Jing and McKeown, 2000; Marsi and Krahmer, 2010). Jing and McKeown (2000) analyzed 30 single document articles and their summaries and characterized the transformations performed on the original text to form a summary. They did not give statistics about the relative frequency of each transformation operation but list "add descriptions or names for people and organizations" and "substitute phrases with more general or specific information" as two of the summarization operations. In a more recent study, Marsi and Krahmer (2010) analyzed the phrase alignment between original spoken news in a Dutch television news program and the subtitles for the same broadcast. They aligned the transcript and the subtitles and analyzed the transformations performed on the phrase level. The authors distinguished five mutually exclusive similarity relations in the corpus: `equals` (the aligned phrases are identical), `restastes` (the aligned phrases convey the same information but with different wording), `specifies` (the subtitle phrase is more specific than the transcript phrase), `generalizes` (the subtitle phrase is more general than the transcript phrase), and `intersects` (the aligned phrases share some informational content, but each also expresses some information not expressed in the other). The second most frequent

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 151–159,
Berlin, Germany, August 11, 2016.

class is `generalizes`[1]. In about 14% of the aligned phrases, the subtitle contained a more general phrase than the original. Only a small percentage of `specifies` pairs is present: in about 3% of the phrases the subtitles were more specific than the transcripts.

Here we present an analysis of generalization operations that occur in abstracts produced for clusters of topically related news articles in Brazilian Portuguese. In the vast majority of cases these require transformations at the phrase level. We observed five types of generalization: interpretation, detail removal, class, role, and whole. Named entity (NE) generalizations, in particular, belong to four categories: detail removal (removing some of the information contained in the original article, similar to compression on the phrase level), role (substituting a reference by name with a reference by the role the entity plays in the described events), class (substituting a reference with a superordinate concept, i.e. "swimmer – "athlete") , and whole (a reference to a member of a group or area is substituted by a reference to the whole, i.e. "Jamaica" – "the Caribbean". In each category, we identified a set of syntactic-semantic operations related to each type of named entities (person, organization, location and sports event). Such operations include substitutions and phrase reductions. Their automation would require the development of capabilities that are not available to current systems.

The remainder of this paper is structured as follows. In Section 2, we introduce our corpus, explaining the manual alignment between the human abstract and the multiple news text inputs, and the pre-processing of such alignments. In Section 3, we describe the analysis of the alignment pairs containing generalization and the categorization of each instance in according to the five-class typology of transformations. Then, in Section 4, our focus relies on the generalization of phrases containing named entities. Specifically, we describe the syntactic and semantic properties of such phrases considering both the different types of generalization and entities. In Section 5, we discuss what we learned and close with discussion of perspectives for automatic summarization.

[1] `Equals` is the most common relation between aligned phrases, accounting for 67% of the alignments

2 The Corpus

We used the CSTNews (Cardoso et al., 2011) corpus of multi-document abstracts and the associated news articles. The corpus comprises 50 clusters of news texts in Brazilian Portuguese from a range of categories: daily news (14), world (14), domestic politics (10), sports (10) , economy (1), and science (1). There are 140 documents in total in the corpus.

Each cluster contains two or three news articles on the same topic, with 42 sentences per cluster on average. There are six manual multi-document abstracts for each cluster. The abstract-writers were instructed to produce abstracts of length equal to 30% of the longest article in the cluster. The resulting abstracts were on average seven sentences (132 words) long. CSTNews has annotated versions of the source texts and summaries in different linguistic levels, e.g., intra- and inter-textual discourse relations, classification of temporal expressions, semantic annotation of nouns and verbs, and subtopic segmentation. The corpus also contains alignments between each human abstract and the source texts at the sentence level. Each sentence in the abstract is associated with all of the sentences in the original articles that support the information expressed in the abstract.

For our work, we use the existing manual annotations, pairing sentences from the abstract with their corresponding sentences in the original article (Camargo et al., 2013). The annotators identified 1,007 alignments, involving 334 summary sentences and 877 document sentences: 99.4% of the summary sentences were aligned to some document sentence and 42.43% of the document sentences were aligned to some some summary sentence.

In addition, for each pair of summary-original sentences, annotators included labels describing the sub-sentential relations between the sentences in the pair. Among other tags, the annotators labeled when a summary sentence contained parts that were more general or more specific than the semantically corresponding part in the document sentence. They however did not mark the exact spans of text involved in the generalization.

The alignment in (1) shows an example of a summary and document sentence that share information and in which one can observe changes in the specificity of reference. The summary sentence has more general content, referring to "many

states" and "the operation" while the document sentence has a list of Brazilian states and the name of the police investigation (shown in bold).

(1) Summary: *Mais de 300 policiais de **[vários estados]** participaram d**[a operação]*** ("More than 300 officers from [many states] were part of the operation").
Document: *Ao menos 300 policiais de **[Amapá, Distrito Federal, Mato Grosso, Acre and Rondônia]** participaram da **[Operação Dominó]*** ("More than 300 officers from [Amazonas, Distrito Federal, Mato Grosso, Acre and Rondônia] were part of the [Operação Dominó"]).

Overall, 13% of summary-document pairs involved a generalization or a specification operation. There are 80 pairs tagged as containing generalization and 47 pairs tagged as containing specification (Camargo et al., 2013). The label describes the change that occurred to transform the document sentence into the summary sentence, i.e. generalization means some information is expressed in more general terms in the summary sentence than it was in the original document sentence.

2.1 Pre-Processing Steps

With the aim categorizing the type of every generalization case in the summary-documents alignments, we performed two manual pre-processing steps: (i) expansion and revision of the alignments with generalization, and (ii) delimitation of the generalization cases and indexing of the textual spans involved in each case.

Abstracts contained both generalizations and specifications of entities. Assuming that the underlying process involved in modifying the reference is the same in both cases, we augment our corpus of generalizations by "inverting" the 47 specification alignments to obtain 47 examples of generalization, as illustrated in (2). The pair is from a news article about the schedule of the Brazilian men's volleyball team. It was originally tagged as specification, since the summary sentence contains more detail than the original; it details that the team aim is to win "the gold medal". We swap the direction of the relation between the sentences and consider the resulting sentences as examples of generalization.

In this way, we obtained a set of 127 pairs of aligned sentences with differences in the specificity of reference. Next, each alignment was manually revised by the first author: 12 of them were excluded because the author did not find clear portions of the summary sentence that generalize information expressed in the original document. An example of sentence that was excluded is given in (3). The final set consists of a total of 115 aligned pairs.

(2) Summary: *O próximo objetivo da seleção é **[a medalha de ouro nos Jogos Pan-Americanos do Rio]*** ("The next goal of the team is [the gold medal in the Pan American Games in Rio]").
Document: *O próximo objetivo é **[os Jogos Pan-Americanos do Rio]*** ("The next goal is [the Pan American Games in Rio]").

(3) Summary: *A pressão argentina continuou no segundo tempo, mas o Brasil fechou a goleada com um gol marcado pela sua dupla de volantes* ("Argentina struggled to make any impact in the 2nd half, but Brazil sealed the victory with a goal made by one of its midfields").
Document: *Os argentinos, com um time repleto de craques favoritos ao título, e com campanha irrepreensível até o momento, pareciam não acreditar no que viam* ("Argentina, a team full of stars and favorite to win, could not believe what was happening").

Next, we carried out an annotation of each pair in order to answer two questions: (1) Which text spans in the respective sentences are involved in the generalization operation? (2) What is the linguistic-level characterization of the spans? The description captured the changes of content from specific to general. Clause generalization was restricted to cases where the summary noun phrase (NP) generalizes a proposition. In order to answer the questions, the spans were marked and labeled according to the corresponding generalization level (C for clause, and P for phrase). If the sentences had more than one case of generalization, they were also numbered according to the order of occurrence in the document, following the notation C/P_NUM. Examples of annotated phrases and clauses are given in (4) and (5), respectively. We extracted a total of 136 pairs of specific-general phrases from the 115 sentence alignments. There are more aligned phrases with

difference in specificity than aligned sentences because some sentence pairs contained more than one case of phrase generalization case.

(4) Document: [O presidente dos EUA, George Bush]**P1**, pediu que o Exército turco busque [uma solução diplomática para a questão]**P2** (“[President of the US, George Bush], asked the Turkish Army to seek [a diplomatic solution to the issue]”)
Summary: [Washington]**P1** e a Comissão Européia também pedem [uma solução diplomática]**P2** (“[Washington] and European Commission also ask for a [diplomatic solution]”)

(5) Document: Na Jamaica, [muitos estocaram comida, água, lanternas e velas]**C** (“In Jamaica, many stock food, water, flashlights and candles”)
Summary: [Muitos moradores e turistas estão se preparando para a passagem do furação. (“[Many locals and tourists prepare for the hurricane]”)

3 Typology of Transformations

Further, we iteratively analyzed the types of the 136 cases of generalization to come up with categories that cover all examples in the corpus. We converged on a classification scheme with five categories: (i) **Interpretation**, i.e., generalization based on sophisticated inferences over the source text and additional information such as transforming ”200 people were injured” to ”the human toll was high”; (ii) **Detail removal**, i.e, generalization by omitting details of a specific textual segment; (iii) **Role**, i.e., replacement of person entities by their title or role; (iv) **Class**, i.e., substitution of a subordinate concept by a superordinate one, and (v) **Whole**, i.e., concepts representing parts are replaced by concepts that indicate the whole. The typology reveals that humans carry out a variety of inferences based on rich world and domain knowledge to produce generic information. Table 1 shows the distribution of the categories divided by clause and phrase levels.

Interpretation is the most frequent category in the corpus (45.6%) and the only one that occurs in both clause and phrase levels. However, 83.8% of the cases (52 out of 62) occur at the clause level and involve propositional generalizations. We show an example in (5). It involves an inference that “stocking food, water, flashlights and candles” is a preparedness activity against hurricane. Detail removal is the second most frequent, with 32 instances (23.5%), followed by Role, with 18 instances (13.2%). The distribution of cases in Class and Whole is quite similar, 13 (9.6%) and 11 (8.1%), respectively. Next, we turn our description to generalizations that occur at the phrase level[2], specifically to those involving named entities.

Category	**Phrase**	**Clause**	Total
Interpretation	10 (7.4)	52 (38.2)	62 (45.6)
Detail removal	32 (23.5)	0	32 (23.5)
Role	18 (13.2)	0	18 (13.2)
Class	13 (9.6)	0	13 (9.6)
Whole	11 (8.1)	0	11 (8.1)
Total	84 (61.8)	52 (38.2)	136 (100)%

Table 1: Number and percentage of the generalizations

Category	**Noun**	**Named entity**	Total
Interpretation	10 (11.9)	0	10 (11.9)
Detail removal	14 (16.7)	18 (21.4)	32 (38.1)
Role	2 (2.4)	16 (19)	18 (21.4)
Class	5 (6)	8 (9.5)	13 (15.5)
Whole	2 (2.4)	6 (10.7)	11 (13.1)
Total	33 (39.3)	51 (60.7)	84 (100)%

Table 2: Number and percentage of general NPs and NEs

4 Named Entity Generalization

We first computed the number of cases that involve named entities or general NPs per category. Table 2 shows the results.

Looking briefly at the 33 common noun pairs, we found that Interpretation tends to be associated with numbers (25%). The substitution of “about 300 buildings” with “many buildings” illustrates this. Interpretation also results from different inferences, e.g., when a cause (e.g., ‘the fog”) is replaced by its effect (e.g., “the bad weather”). The Role case where “the 16 children and 14 adults” was replaced with “the 30 hostages” is the only one involving generation of a numeric expression. Detail removal occurs by deleting noun adjuncts (shown in italics) (e.g., “a *university* campus”) or complements (e.g., “the inspection *of income tax declarations*”).

Studying in detail the 51 generalizations involving NEs, we found four types of NEs: 26 persons (51%), 16 organizations (31.4%), 7 locations (13.7%), and 2 sports events (2%). We also identified sub-categories of generalization for three en-

[2]Appendix 1 (Table 4) provides examples of phrase generalizations.

Entity	Category	Sub-category	Document Phrase	Summary Phrase
Event (2)	Class (2)	–	Name (2)	Noun+Post-mod (2)
Location (7)	Whole (4)	Island-to-region (1) City-to-state (1) City-to-country (2)	Name (4)	Name (4)
	Detail removal (3)	–	Pre-mod+Name (1) Name (2)	Noun (3)
Organization (16)	Class (6)	–	Name (6)	Noun (6)
	Detail removal (6)	–	Name (4) Name+Post-mod (2)	Noun (4) Acronym (2)
	Whole (4)	Member-to-organization (4)	Name (2) Name (2)	Noun (2) Noun (1), name (1)
People (26)	Role (16)	–	Pre-mod+Full Name (6) First Name (4) Last name (3) Pre-mod+First name (1) Pre-mod+Last name (1) Acronym (1)	Noun (16)
	Detail removal (9)	–	Pre-mod+Full name (5)	Noun (3), First name (2)
			Full name (3)	First name (3)
			Pre-mod+Last name (1)	Noun (1)
	Whole (1)	Person-to-place (1)	Pre-mod+Full name (1)	Noun (1)

Table 3: Semantic and syntactic properties of named entity phrase generalization

tity types and some syntactic patterns in the transitions, related to the type of the phrase head and the occurrence of pre- and post-modifiers. The results are shown in Table 3. The numbers in parenthesis show how many times the given category and syntactic form have occurred in the pairs.

The sports event generalizations consist in substituting the multi-word expression (MWE) phrase "the American Cup" with two different general NPs: "the continental competition" and "the oldest soccer tournament". These are the only instances where the summary phrases include modification. Thus, both general mentions put the referent in a class and provide further details about it as well.

According to Table 3, there are three types of Whole generalizations for locations that solely involve names: (i) `island-to-region`, such as the replacement of "Haiti" and "Dominican Republic" with "the Caribbean"; (ii) `city-to-state`, such as "Maceió, which was substituted by "Alagoas", and (iii) `city-to-country`, such as the replacement of "Boston" with "United States". There is also one particular type of Detail removal by deleting names from phrases of the form pre-modifier + name (e.g., "the capital Kingston") to produce mentions whose head was the modifying noun of the specific phrase ("the capital"). Location names, specifically MWEs (e.g., "International Airport of São Paulo") that are made up of a place (possibly a MWE itself, such as "São Paulo") and additional information (e.g., "international"), are also replaced with common nouns ("the airport"). The replacement of such proper names with common nouns result from removal of all the details about the referent description.

Organization names are mostly generalized by means of common nouns that express class or whole. The substitution of "Brazil" with "the country" illustrates the Class category. The Whole generalization occurs through `member-to-organization` substitution, i.e., the replacement of "the Military Police Shock Troop" with "the police" illustrates this. The only case of name generalization is the substitution of "the Archdiocese of Los Angeles" with "the Catholic Church". There are also cases where names followed by acronyms in parenthesis, such as "National Institute of Social Security (INSS)", are reduced to the acronym only.

It can be seen that document mentions to people have different head types: full name, first name, last name, and acronym. With the exception of acronyms, the heads usually occur with two types of pre-modifiers (shown in italics): titles (e.g.,"*president* of the Senate, Renan Calheiros") and roles (e.g., "the *goalkeeper* Vieri"). In general, the document mentions are commonly replaced with common nouns only. The substitution of the first name "João Pedro" with "the senator" illustrates this. The summary writers also chose the modifying noun (shown in italics) from phrases of the form pre-modifier + name (e.g., "the *goalkeeper* Vieri") for generalization, deleting the last or fulll name (e.g., "the goalkeeper").

The reduction of full name by deleting surname (shown in italics) (e.g., "Renan *Calheiros*"), yielding phrases containing first name only (e.g., "Renan"), is another common type of operation. The case that belong to the Whole category is the only one involving two different types of named entity. In particular, "President of the United States, George Bush" was substituted by "Washington", in a `person-to-place` operation.

5 Discussion

This study provides an initial characterization for phrase generalizations that arise in summarization. It is evident that our results should be validated on a larger sample of summarization data. Nevertheless our findings can be seen as a good start for understanding the phenomenon. One of the practical outcomes from our work is the generalization typology which can be applied for the analysis of other data.

Interpretation is the most common category, resulting from inferences over propositions and covering a variety of operations. Its automatic treatment would be a major endeavor in natural language processing research because it is at the intersection of semantic interpretation and text generation.

Another challenge for summarization systems is how to deal with mentions of numbers, which form a special class of the interpretation transformation. We found that references to date, time, and general quantities accounted for 25% (8 out of 33 instances) of common noun phrase alignments in our corpus. Only in one case the numeric expression was transformed in an alternative numeric expression. All other phrases involving numbers were lexicalized alternatively. Then the task of a system would be to identify which references to numbers should be generalized and how to generate the generalization of numbers.

In our study, 61% of the generalizations involve operations over specific mentions to named entities. These have been studied computationally in the past, to predict the appropriate form of the name in references to people (Siddharthan et al., 2011) and to exploit the person name repetition in the summary to find the salience of entities (Dunietz and Gillick, 2014). Neither of these prior studies analyzed reference to named entities by common noun, which we provided in the analysis of our data, nor do they look at non-person references. In fact, substituting names with generic nouns was the most common operation in our data and it calls for the development of new capabilities, both to decide which entities should be mentioned generically and how to generate the reference itself.

Moreover, specific mentions to sports events, locations and organizations do not include modification in 88% (22 out of 25) of the pairs. Specific mentions to people have an accompanying description in around half the cases (57.6%). The occurrence of a pre-modifying word that identify the person's title or role provides more details about the referent. Thus, such mentions have a higher level of specificity than other with name only. Moreover, only few generic phrases contain a name, and, when it occurs, the names have particular types, e.g., first name in the case of people, and acronyms for organization.

On the operations concerning named entities, we provide some insights for substitution and reduction approaches to obtain general phrases.

Substitution is the most common operation (76.5%) (out of 51 cases), and its automatic process would require structured knowledge that includes at least three relationships: (i) **is-a** to express the rough notion of "a kind of", (ii) **part-whole** to express `island-to-region`, `city-to-state`, `city-to-country`, `member-to-organization`, and `person-to-place`, and (iii) **instance-role**, for entities of the person category. Since such knowledge is very particular to some domains, specially global and local sports, politics, and geography, we believe that it would possible to model it in handcrafted lexicons. It could also be derived automatically for some types of reference (McKinlay and Markert, 2011; Mitchell et al., 2015). In addition, modules to decide when substitution is necessary or appropriate would be needed.

Phrase reduction (i.e., deletion of words or phrases) occurs in 23.5% of the cases (out of 51). Although it is less frequent, detail removal include cases where specific phrases could be automatically converted into general in a more feasible way. This observation is based on the fact that summary phrases are made up of linguistic material that came from the document phrases. Thus, we can conceive phrase reduction as a similar task to sentence compression, where the oper-

ations are learned by analyzing pair of sentences, one from the source text, and other from human-written abstracts such that they both have the same content. Specifically, 4 reduction rules could be defined: (i) removing pre-modifier from phrase of the form modifier + location name, yielding a common noun mention; (ii) removing name from phrase of the form organization name + parenthetical acronym, generating an mention with acronym only; (iii) deleting name from phrase of the form title/role + person name, producing a common noun mention, and (iv) removing surname from person full name, generating a first name mention.

We may also contribute for generating references, since referring expressions in extracts can be problematic because the sentences compiled from different documents might contain too little, too much, or repeated information about the referent. Our results show that 76.5% of the 51 generalizations with named entities (e.g., "the coach Bernardinho") are made solely with a common noun phrase (without the inclusion of the entity's name) (e.g., "the coach"), and thus a task to be considered is the generation of common noun references to named entities. Such generation would allow the production of a more natural summary.

We are aware of full coreference resolution is a very difficult problem and there are no systems that can reliably perform it on free texts. But we believe that the availability of cross-document information can facilitate the resolution of common noun phrases. This assumption is built on the fact that most common nouns in summary phrases were contained in the input texts. For example, the head of the summary NP "the coach", which generalizes the name "Bernardinho", is contained in a different sentence of the same input, as part of the mention "the coach Bernardinho". This means that lexical overlap would indicate that these three NPs refer to the same entity. Common noun generation would increase the genericity level of summaries, and avoid the repetition of forms produced by some rewriting methods (Siddharthan et al., 2011).

6 Future work

Our research both provides a preliminary characterization of generalization in document-summary alignments and a discussion of some insights for Natural Language Processing. For future work, we plan to increase the sample of specific-generic pairs by aligning the five new abstracts recently added to each cluster of CSTNews in order to validate our results. We could repeat the manual alignment or use automatic methods (Agostini et al., 2014). To identify the categories, we intend to carry out a manual annotation with multiple judges.

Moreover, we have been performing a manual annotation of coreference chains that consist of all the mentions of an entity in abstracts with different lengths in two languages, Portuguese and English. Our goal is to explore human preferences in mention realization, and possible differences across languages. We also aim at exploring whether the abstract length has influence on the syntactic forms and sequences of mentions, and on the amount of information included in the mentions.

Acknowledgment: We thank the State of São Paulo Research Foundation (FAPESP) (#2015/01450-5) for the financial support.

References

Verônica Agostini, Roque E.L. Condori, and Thiago A. S. Pardo. 2014. Automatic alignment of news texts and their multi-document summaries: Comparison among methods. In *Proceedings of the 11st International Conference on Computational Processing of Portuguese*, pages 286–291, São Carlos, SP, Brazil.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th ACL/HLT - Volume 1*, pages 481–490. Association for Computational Linguistics.

Renata T. Camargo, Verônica Agostini, Ariani Di-Felippo, and Thiago A. S. Pardo. 2013. Manual typification of source texts and multi-document summaries alignments. *Procedia Social and Behavioral Sciences*, 95:498–506.

Paula C. F. Cardoso, Erick G. Maziero, Maria Lucia R. Castro Jorge, Eloize M. R. Seno, Ariani Di-Felippo, Lucia Helena M. Rino, Maria das Graas V. Nunes, and Thiago Pardo. 2011. Cstnews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, Cuiabá, MT, Brazil.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 137–144.

Jesse Dunietz and Dan Gillick. 2014. A new entity salience task with millions of training examples. In *Proceedings of the European Association for Computational Linguistics*, pages 2282–2287.

Michel Galley and Kathleen McKeown. 2007. Lexicalized markov grammars for sentence compression. In *HLT-NAACL*, pages 180–187.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18.

Hongyan Jing and Kathleen R. McKeown. 2000. The decomposition of human-written summary sentence. In *Proceedings of the 1st NAACL Conference*, pages 178–185, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence*, pages 703–710, Austin, Texas, USA.

Alex Kulesza and Ben Taskar. 2011. Learning determinantal point processes. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 419–427.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 510–520.

Erwin Marsi and Emiel Krahmer. 2010. On the limits of sentence compression by deletion. In Erwin Marsi and M. Theune, editors, *Empirical Methods in Natural Language Generation*, pages 45–66. Springer-Verlag, Berlin, Heidelberg.

Ryan T McDonald. 2006. Discriminative sentence compression with soft syntactic evidence. In *EACL*.

Andrew McKinlay and Katja Markert. 2011. Modelling entity instantiations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 268–274.

Tom M. Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapa Nakashole, Emmanouil Antonios Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. 2015. Never-ending learning. In *Proceedings of the European Association for Computational Linguistics*, pages 2302–2310.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.

Stefan Riezler, Tracy H King, Richard Crouch, and Annie Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *Proceedings of the NAACL/HLT'03*, pages 118–125.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Jenine Turner and Eugene Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 290–297.

.

Types	Specific segment	Generic segment
Interpret.	cerca de 22 pessoas	a maioria das vítimas
	("about 22 of the victims")	*("the most victims")*
	os primeiros 4 minutos de jogo	the fourth minute of the match
	("the fourth minute of the match")	*("the begnning of the match")*
	casas e viadutos destruídos	grandes danos materiais
	("destroyed houses and viaducts")	*("great damage")*
	dois terços das autuações de contribuintes	as irregularidades mais comuns
	("two-thirds of the taxpayers' infractions")	*("the most common infractions")*
	(às) 11h40	(por) a manhã
	("at 11h40")	*("(in) the morning ")*
	cerca de 300 edifícios	vários edifícios
	("about 300 buildings")	*("several buildings")*
	a polícia *("the police")*	o governo *("the government")*
	o nevoeiro *("the fog")*	o mau tempo *("the bad weather")*
	o ajuizamento de uma ação civil pública	medidas necessárias
	("the filing of a public civil action")	*("necessary measures")*
Detail removal	um campus universitário *(an university campus)*	um campus *("a campus)*
	o goleiro Vieri *("the goalkeeper Vieri")*	o goleiro *("the goalkeeper")*
	as Ilhas Cayman *("the Cayman Islands")*	as ilhas *("the islands")*
	quase metade dos voos *("almost half of the flights")*	metade dos voos *("half of the flights")*
	a Operação Farrapos, da Polícia Federal	a operação
	("the Federal Police's "Operation Farrapos"")	*("the operation")*
	Instituto Nacional do Seguro Social (INSS)	INSS
	("National Institute of Social Security (INSS)")	*INSS*
	a pista principal do aeroporto *("the main runway")*	uma das pistas *("one of the runways")*
	a medalha de ouro nos Jogos Pan-Americanos	os Jogos Pan-Americanos
	("the gold medal in the Pan-American Games")	*("the Pan American Games")*
	o Aeroporto Internacional de Guarulhos	o Aeroporto de Guarulhos
	("the International Airpot of Guarulhos")	*("the Guarulhos Airport")*
	a capital Kingston *("the capital Kingston")*	a capital *("the capital")*
	falência de órgãos secundária à insuficiência cardíaca	insuficiência cardíaca
	("organs failure secondary to heart disease")	*("heart failure")*
Role	Peterka (*Roberto Peterka)	um perito aposentado *("a retired expert")*
	o advogado das supostas vítimas, R. Boucher	os advogados
	("the lawyer of the alleged victims, Boucher")	*("the lawyers")*
	as 16 crianças e 14 adultos	as 30 vítimas
	("the 14 children and 14 adults")	*("the 30 hostages")*
	uma quadrilha de altos funcionários públicos	pessoas suspeitas
	("a group of high-level public officials (accused of fraud)"	*("suspicious people")*
Class	os Estados Unidos *("the United States")*	o país *("the country")*
	o revólver *("the revolver/gun")*	as armas *("the weapons")*
	Abadia (*Juan Carlos Ramírez Abadía)	o colombiano *("the Colombian")*
	a queda (do avião) *("the crash")*	o acidente *("the accident")*
	a Schincariol *("the Schincariol")*	a empresa *("the company")*
Whole	Maceió (*capital of Alagoas)	Alagoas (*Brazilian state)
	a Arquidiocese de Los Angeles	a Igreja Católica
	("The Archdiocese of Los Angeles")	*("The Catholic Church")*
	o Depart. de Investigações sobre Crime Organizado	a polícia
	("the State Department of Criminal Investigation")	*("the police")*

Table 4: Examples of phrase-level generalization from the CSTNews corpus (Appendix 1)