# Source-summary Entity Aggregation in Abstractive Summarization

**José Ángel González**
Valencian Research Institute
for Artificial Intelligence, Valencia
jogonba2@dsic.upv.es

**Annie Louis**
Google Research, London
annielouis@google.com

**Jackie C. K. Cheung**
McGill University/MILA, Montreal
Canada CIFAR AI Chair
jcheung@cs.mcgill.ca

## Abstract

In a text, entities mentioned earlier can be referred to in later discourse by a more general description. For example, *Celine Dion* and *Justin Bieber* can be referred to by *Canadian singers* or *celebrities*. In this work, we study this phenomenon in the context of summarization, where entities from a source text are generalized in the summary. We call such instances *source-summary entity aggregations*. We categorize these aggregations into two types and analyze them in the CNN/DAILYMAIL corpus, showing that they are reasonably frequent. We then examine how well three state-of-the-art summarization systems can generate such aggregations within summaries. We also develop techniques to encourage them to generate more aggregations. Our results show that there is significant room for improvement in producing semantically correct aggregations.

## 1 Introduction

The quality of abstractive summarization systems has improved substantially in the past few years. An important next research question is to better understand the specific linguistic and semantic operations which can lead to high-quality abstractive text. In this work, we focus on how entities can be referred to in summaries, especially with an expression more general than in the source. For example, Table 1 demonstrates how three comic book characters mentioned in the source document are aggregated in a reference summary by the expression, "three of the most well-known comic book characters of all time". Such referring expressions are interesting for abstractive summarization, since they are *novel summary n-grams* that result from semantic inference from the source.

There are few existing studies about the semantics of text generated by current abstractive systems. Some have focused on summary n-grams that are not found in the source text (Kryściński et al., 2018; Song et al., 2020), and others that look at problems resulting from undesirable summary content, e.g., hallucinations (Maynez et al., 2020; Kryściński et al., 2020).

| | |
|---|---|
| **Document** | (CNN) Comic books of the past few years have seen a lot of changes (a female Thor, anyone?) but not quite so many at one time. Three major characters – Superman, Wonder Woman (both of DC Comics, a Time Warner company, like CNN) and Archie Andrews – came out with new looks (...) |
| **Summary** | Superman, Wonder Woman and Archie all debuted new looks Thursday. Three of the most well-known comic book characters of all time look radically different. |

Table 1: An example of source-summary entity aggregation. The aggregation "*Three of the most well-known comic book characters of all time*" is used in the summary to aggregate the entities "*Superman*", "*Wonder Woman*", and "*Archie Andrews*" named in the document.

We focus on a specific semantic operation that summary writers can perform in order to change the level of detail in a summary: the semantic aggregation of named entities in context, as in Table 1. We estimate the prevalence of such aggregations in the CNN/DAILYMAIL corpus (Hermann et al., 2015). We also categorize the aggregations that we find into those (i) where the models can *copy* the aggregations from the source document, and (ii) those cases where the models are required to generate *novel* aggregations not found in the source.

We then explore how well existing systems can generate *copy* and *novel* aggregations that match those found in reference summaries. Specifically, given a document, the models must generate a summary, and the aggregations within the generated summary are evaluated against the aggregations in the reference summary.

We evaluate three state-of-the-art Transformer-based abstractive summarization systems: BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020),



*Proceedings of the 29th International Conference on Computational Linguistics*, pages 6019–6034
October 12–17, 2022.

and $T_5$ (Raffel et al., 2020). The experimental results show that the task is hard, especially for generating novel aggregations. We also explore how to fine-tune BART (Lewis et al., 2020) to generate more summary-worthy aggregations without compromising the overall summary quality. The performance of all the summarization models is still far below the upper bounds posed by our oracles, showing that there is room for improvement.

## 2 Related work

Semantic generalization in automatic summarization is receiving increasing interest, both in terms of the data and the models. In Belkebir and Guessoum (2016), the authors fuse concepts within sentences using hypernymy relations from taxonomies such as WordNet. Kouris et al. (2019) focuses on abstracting single concepts. Roughly, they train an encoder-decoder architecture on documents where single nouns are replaced with hypernyms, to produce more general summaries. Contrary to these approaches based on taxonomies, Kryściński et al. (2018) use a mixed objective for training encoder-decoder architectures to encourage abstraction in summaries. The level of abstraction was defined in terms of novel n-grams.

The surface-level novel n-grams definition of abstractiveness has also been used in recent summarization datasets (Grusky et al., 2018; Narayan et al., 2018). This approximation is convenient for generation since it measures any kind of rewriting. However, being able to explicitly measure different types of abstraction is important for tracking progress. Our work is based on this motivation. The closest idea towards entity aggregation is Jumel et al. (2020) and we draw heavily from their work. They introduce a dataset and task (TESA) which consists of producing a non-enumerating noun phrase ('former US presidents') that aggregates a set of entities ('Clinton', 'Bush') in a textual context (a New York Times article). Their data was collected using crowd annotators and does not specifically focus on any task. Our work explores entity aggregations in the context of abstractive summarization.

Our task can be seen as a referring expression generation problem (Stone, 2000; Krahmer and van Deemter, 2012) where a general phrase in the summary stands in for a set of entities in the source. The task is also related to multi-antecedent coreference resolution and split-antecedent anaphora (Yu et al., 2020; Burga et al., 2016; Vala et al., 2016), but most resources and approaches here are aimed at pronominal coreference. Some studies address entity-driven summarization (Zhou et al., 2021; Sharma et al., 2019), with the aim of focusing the summaries on the most salient entities. Differently to our work, this work does not focus explicitly on the aggregation expressions in the summaries.

Currently, abstractive summarization systems have vastly improved generation capabilities, achieved by using pre-trained Transformers (Vaswani et al., 2017) as a backbone. Among the state of the art of text summarization benchmarks, BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), PROPHETNET (Qi et al., 2020) and $T_5$ (Raffel et al., 2020) stand out in terms of ROUGE. Some challenges with these models, such as mitigating hallucinations or ensuring factual consistency, are of great interest. Different methods have been proposed here (Zhao et al., 2020; Nan et al., 2021b). Of these Narayan et al. (2021); Nan et al. (2021a) incorporate text planning using a sequence of entities to prompt faithful generation. We use such ideas to encourage systems to produce more aggregations by jointly training the models to identify summary-worthy aggregations while learning to generate summaries.

## 3 Corpus Study

### 3.1 Defining source-summary entity aggregations

We categorize source-summary entity aggregation into two types: a) **copy aggregations** found in the source and b) **novel aggregations** that must be generated. In copy aggregations, an aggregating expression for entities is present in the source and the same expression is also in the summary. In the novel case, entities from the source are aggregated by a new expression not found in the source. Table 2 shows examples of the two types.

This distinction is useful for summarization systems since a system will understandably find it easier to copy an expression from the source. Generating novel aggregations is likely more difficult, in theory requiring deeper semantic understanding of the source content.

### 3.2 Data Source

One would like to understand how often aggregations are used in human summaries and also how well systems currently handle them. However,

| Document | Summary | Type |
|---|---|---|
| Through the process of elimination (...), our guess as to which five states. White will play on the brief acoustic run: South and North Dakota, Wyoming, Vermont and ... Puerto Rico? | Jack White taking a hiatus from touring after brief acoustic jaunt . He'll play five states he has yet to get to, charge just $3 . Places and times of shows are currently a mystery. | *Copy* Aggregation |
| Camuti and Rakes were longtime business associates, and Camuti allegedly poisoned Rakes at a time when Camuti owed money to Rakes. (...) | William Camuti, 69, is charged with attempted murder and misleading police. Camuti and victim Stephen Rakes were longtime business associates. | |
| Sometime in the not-too-distant future, Kanye West can once again (...) The rapper and his reality TV star girlfriend, Kim Kardashian, are having a girl. | News comes the same day Kardashian has a baby shower. The couple has been dating since last year . The baby is due next month. | *Novel* Aggregation |
| After a youth rally in the Bahamas National Stadium Monday, Harry travels to Jamaica and then on to Brazil to complete his 10-day tour (...) | The Bahamas is Prince Harry's second stop in a 10-day Caribbean tour . The 27-year-old prince is celebrating the Diamond Jubilee of Queen Elizabeth II . Jamaica and Brazil are Harry's next two destinations. | |

Table 2: Examples from the CNN/DAILYMAIL for the different cases of source-summary entity aggregation. Aggregations are in green and entities are in blue.

doing so requires a dataset with entities marked together with their aggregations, also aligned between source and summary texts. Such annotations are costly to produce in practice. So, we approximate counts using heuristics which we will outline in this section.

We use the TESA (Jumel et al., 2020) dataset, where sets of entities are paired with human-written aggregation expressions. We briefly describe this dataset and then how we designed our heuristics.

The TESA corpus comprises 1,718 'aggregatable instances', each one consisting of (a) a set of named entities of the same type (person, location, and organization), (b) a context (excerpt) from an article in the NEW YORK TIMES corpus (Sandhaus, Evan, 2008) involving the entities, (c) background knowledge of the entities from Wikipedia, and (d) at least one human-written aggregation. An example of an aggregatable instance is shown in Table 3.

| | |
|---|---|
| **Background** | **Microsoft**: Microsoft Corporation is an American multinational technology company (...)<br>**Sony**: Sony Corporation is a Japanese multinational conglomerate corporation (...) |
| **Context** | Battleground For Consoles Moves Online: Over all, though, it is Microsoft that has had the steeper mountain to climb. In the last generation of video game consoles, Sony had a roughly 60 percent market share, compared to 20 percent for each Microsoft and Nintendo. |
| **Aggregations** | technology companies, multinational corporations |

Table 3: Example of an aggregatable instance from TESA. The entities of the set are in blue and the annotated aggregations are in green.

Table 4 shows several basic statistics of TESA. Table 9 in Appendix A shows some of the most frequent entity aggregations. Note that these examples are not specific to summarization, and they show entity aggregation in the context of a news article. In addition, the corpus only contains named entities which are persons, organizations, or locations. So, this set is only a small subset of possible source-summary relations. However, given the difficulty of defining and identifying such aggregations, we similarly limit our work to the same types of named entities.

| | Entity Sets | Entities | Aggregations |
|---|---|---|---|
| Person | 941 (801) | 2228 (1201) | 2900 (951) |
| Location | 629 (412) | 1606 (278) | 2041 (505) |
| Organization | 148 (123) | 310 (196) | 456 (239) |

Table 4: Statistics of each type in the TESA corpus. We indicate the total count of occurrences, and in parentheses the count of unique occurrences.

### 3.3 Prevalence in the CNN/DailyMail corpus

To determine the prevalence of source-summary entity aggregations in the CNN/DAILYMAIL corpus, we computed the percentage of document-summary pairs containing each type of aggregation. We do this heuristically due to the difficulty of labeling. This process consists of the following steps:

1. Identify sets of entities in a source document that could be aggregated.

2. Identify the possible aggregations in the source and summary documents that could be matched with each source entity set.

#### 3.3.1 Detecting aggregatable entity tuples

We used four lexico-syntactic patterns to detect tuples of entities in source documents which could potentially be aggregated. Entities[1] in a tuple must

[1]We used the Spacy pipeline (Montani et al., 2022) with RoBERTa-base (Liu et al., 2019) to extract named entities and noun phrases.

be of the same entity type (person, location, or organization). These patterns are:

- **Coordinating conjunctions**: list of entities separated by , or ; ending with a conjunction (*and/or*), e.g., *John, Peter, and Mary*.
- **In sentence**: entities mentioned at any position within the same sentence, excluding the **Coordinating conjunctions** pattern, e.g., *Today John meets Mary*.
- **Contiguous sentences**: entities used to begin contiguous sentences, e.g., *John went to the beach. Mary went to the mountain.*
- **Shared nouns**: entities mentioned at any position that are preceded by the same singular noun phrase, e.g., *Rock climber John prefers a change of scenery. In contrast, rock climber Mary prefers the mountain.*

This step gives us aggregatable entity tuples from a source document.

#### 3.3.2 Identifying possible aggregations

This step identifies likely aggregation expressions. First, we find likely expressions for these entity tuples in the source, where possible. For that, we use a collection of heuristics to pick candidate expressions within close proximity to the entities. We will then explain how we align these aggregations to those in the summary (to identify if they are novel or copied from the source).

The following heuristics identify aggregations in close proximity to entity tuples *in the source*:

- **Previous sentence**: a noun phrase in the sentence that precedes the span of sentences containing the entities, e.g., *The rock climbers are traveling by the world. Some of them are John and Mary.*
- **In span**: a noun phrase in the span of sentences containing the entities, e.g., *John went to the beach, he is one of the traveling rock climbers. Mary went to the mountain.*
- **Next sentence**: a noun phrase in the sentence that follows the span of sentences containing the entities, e.g., *Today John meets Mary. Both rock climbers will have a virtual meeting.*
- **Preceding a list**: a noun phrase which introduces a list, with phrases such as (*like*, *such as*, *including*), whitespaces, commas, semicolons, and '—', e.g., *Young rock climbers such as John and Mary are traveling today*. This pattern can only be used to get aggregations of entities obtained by **Coordinating Conjunctions**.
- **Preceding entities**: the plural form of a singular noun phrase that precedes all the entities, e.g., *Rock climber John prefers a change of scenery. In contrast, rock climber Mary prefers the mountain.* This pattern can only be used with **Shared nouns**.

Note that these candidate aggregations may be noisy. So, we only select a noun phrase if it contains an aggregation expression from the TESA corpus. For example, 'young rock climbers' would be selected if any of 'young rock climbers', 'rock climbers' or 'climbers' is present in TESA.

At the end of this step, we have tuples of entities in a source document. Each tuple may be mapped to a list of possible aggregation expressions in the source or no aggregation at all. We show one example of an entity tuple and an associated identified aggregation from the CNN/DAILYMAIL: "*He is the first Western leader to visit one of the three worst affected west African countries - Liberia, Sierra Leone and Guinea.*". Table 10 in Appendix A shows more examples.

We now use these sets to identify copy aggregations and novel aggregations *in summaries of these documents*.

**Copy aggregations:** are source aggregations which are also present in the summary. When an entity tuple does not have a source aggregation from the previous step, we drop it. For the remaining tuples, we check whether one of its aggregations is present in the summary. That entity tuple-aggregation pair is a copy aggregation.

**Novel aggregations:** Here, the aggregation must appear in the summary and not in the source. So the extracted entity tuples and their 'source' aggregations are used differently here.

We create an overall map to be used across the whole set of documents. For each aggregation expression, we aim to create a list of entities, any subset of which can be aggregated by that expression. We do this by merging the aggregation expressions and entity tuples identified in CNN/DAILYMAIL using the heuristics in this section. For example, suppose the corpus contains two documents A and B.

Document A has the source tuple-aggregation pair {'Biden', 'Harris'} → 'politicians' and document B contains {'Modi', 'Johnson'} → 'politicians', then the overall map will contain 'politicians' → {'Bush', 'Clinton', 'Johnson', 'Modi'}. Note that this step merges the entities for the same expression across all of the CNN/DailyMail corpus. To increase the coverage of this overall map, we also add aggregation-entity pairs from the TESA corpus. i.e. If TESA contains the annotation {'Modi', 'Johnson'} → 'prime ministers', the entry 'prime ministers' → {'Johnson', 'Modi'} is also added to our table.

This map is now a broad list of aggregations and possible candidate entities. To identify novel aggregations, we find those cases where an aggregation expression from the table is in the summary but not in the source, and any subset of entities from its entity list is present in the source. Note that this subset may be {'Modi', 'Biden'} which matches a new document C now.

Because our heuristics differ in terms of precision and recall, we computed *low* and *high* estimates of the percentage of documents containing each case. For the low estimates, we only used **Coordinating conjunctions+Preceding a list** since it showed almost 100% precision in our preliminary evaluation (but a very low recall). For the high estimates, all the heuristics are used.

We found that up to 15% of all document-summary pairs in the CNN/DAILYMAIL corpus could contain some type of source-summary aggregation, being thus a reasonably frequent phenomenon even in such an extractive dataset. Following the low estimate, *novel* aggregations are more frequent than *copy* aggregations from the source (1.13% vs. 0.64%). However, following the high estimate, it seems that *copy* aggregations are more frequent (10.95% vs. 4.99%).

Note that to reduce noise, we filtered our expressions using the manual expressions from TESA which covers New York Times articles mostly from an earlier time period, and only entities which are salient and of fixed named entity types. Consequently, we are likely underestimating the prevalence of source-summary entity aggregation in the CNN/DAILYMAIL corpus.

## 4 Experimental setup

Our experiments aim to evaluate the capabilities of state-of-the-art summarization models to generate summary-worthy aggregations. In this section, we describe the task, our oracles, models and evaluation measures.

### 4.1 Task definition

Given a document, the models must generate a summary, and the aggregations within the generated summary must match or be close to those in the reference summary.

We built two test sets from the development and test partitions of the CNN/DAILYMAIL corpus: the COPY and NOVEL sets. We make this distinction to independently evaluate *copy* and *novel* aggregations. To create these sets, we use a broad heuristic compared to our corpus study. Here we gather all noun phrases in the summary and check if their span contains a TESA aggregation. If so, we call it an aggregation and check it against the source to separate into copy and novel sets. Note that here we do not obtain an alignment with entity tuples in the source as in our corpus study. Such alignments would have lower coverage and greater noise, hence we opt for this simple heuristic here.

Therefore each sample in these sets is a triple (D, S, $\mathcal{A}$), where D is a document, S its summary, and $\mathcal{A}$ is the set of aggregation expressions in S. All our samples have non-empty $\mathcal{A}$ sets. In the COPY set, $\mathcal{A}$ only contains aggregations that also appear verbatim in the document. In the NOVEL set, the aggregations $\mathcal{A}$ do not appear verbatim in the document.

Basic statistics of both test sets are shown in Table 5. Table 6 shows the most frequent aggregations. We show some examples of source and references in Table 12 of Appendix A.

| | COPY | NOVEL |
|---|---|---|
| Examples | 4156 | 2905 |
| Avg words (document) | 754.74 (351.70) | 750.72 (365.52) |
| Avg words (summary) | 61.44 (31.63) | 61.53 (39.22) |
| Avg words (aggregations) | 1.78 (0.82) | 1.12 (0.38) |
| Unique aggregations | 1744 | 2120 |
| Avg of aggregations | 1.22 (0.51) | 1.12 (0.38) |
| %CNN/DAILYMAIL val+test | 16.71% | 11.69% |

Table 5: Statistics of the COPY and NOVEL sets. Standard deviations are in parentheses.

### 4.2 Evaluation measures

We evaluate the performance of the models from two points of view: at the aggregation level (how well do the aggregations of the generated summaries match those of the reference summaries?) and at the summary level (how good are the generated summaries compared to the references?).

| | | |
|---|---|---|
| COPY | Person | women (257), friends (173), family (160), men (136), his family (113) |
| | Location | countries (17), cities (14), locations (6), communities (6), states (4) |
| | Org. | schools (33), companies (17), businesses (15), teams (13), groups (11) |
| NOVEL | Person | the pair (117), his family (49), the couple (29), the men (24), officials (20) |
| | Location | other countries (4), countries (4), other cities (3), communities (3), cities(3) |
| | Org. | schools (6), the two teams (4), both teams (4), other teams (3), record labels (3) |

Table 6: The five most frequent aggregations per entity type in COPY and NOVEL test sets. We indicate in parentheses the count of occurrences.

For the aggregation-level evaluation, we consider three metrics. Two are based on overlaps (**Aggregation exact match** and **Token exact match**) and the third relies on similarities among aggregations (**Aggregation relaxed match**).

**Aggregation exact match**: Let $\mathcal{A}_{ref}$ and $\mathcal{A}_{gen}$ be the aggregations in the reference and the generated summary respectively. Precision is defined as $P = \frac{|\mathcal{A}_{ref} \cap \mathcal{A}_{gen}|}{|\mathcal{A}_{gen}|}$ and recall as $R = \frac{|\mathcal{A}_{ref} \cap \mathcal{A}_{gen}|}{|\mathcal{A}_{ref}|}$.

**Token exact match**: Let $\mathcal{T}_{ref}$ and $\mathcal{T}_{gen}$ be the sets of words in the aggregations of the reference and the generated summary respectively. Precision is defined as $P = \frac{|\mathcal{T}_{ref} \cap \mathcal{T}_{gen}|}{|\mathcal{T}_{gen}|}$ and recall as $R = \frac{|\mathcal{T}_{ref} \cap \mathcal{T}_{gen}|}{|\mathcal{T}_{ref}|}$. This score does not constrain the expression phrases to match exactly.

**Aggregation relaxed match**: This variant also measures matches when the generated aggregations do not have word overlap with the reference aggregations, e.g., $\mathcal{A}_{ref} = \{news\ websites\}$ and $\mathcal{A}_{gen} = \{online\ newspapers\}$. In those cases, the two previous measures are too restrictive. So, we propose a measure based on similarities among the aggregations, computed by using BERTSCORE (BS) (Zhang et al., 2019). The precision and recall of this measure are defined as follows:

$$P = \frac{1}{|\mathcal{A}_{gen}|} \sum_{a_{gen} \in \mathcal{A}_{gen}} \max_{a_{ref} \in \mathcal{A}_{ref}} \text{BS}(a_{gen}, a_{ref}) \quad (1)$$

$$R = \frac{1}{|\mathcal{A}_{ref}|} \sum_{a_{ref} \in \mathcal{A}_{ref}} \max_{a_{gen} \in \mathcal{A}_{gen}} \text{BS}(a_{gen}, a_{ref}) \quad (2)$$

For the summary-level evaluation, we report both ROUGE-$F_1$ scores (Lin, 2004) and BERTSCORE to assess the generated summaries. We do not perform human evaluations of content quality because we only want to check that it has not dropped drastically.

### 4.3 Models

We used three state-of-the-art abstractive summarization models as the main systems: BART[2] (Lewis et al., 2020), PEGASUS[3] (Zhang et al., 2020) and $T_5$[4] (Raffel et al., 2020). All these systems are fine-tuned using the training set of the CNN/DAILYMAIL, and evaluated on the COPY and NOVEL sets. All models were implemented using HuggingFace Transformers (Wolf et al., 2020).

Early findings showed that BART generates better aggregations than the other two. So, we also explored how to fine-tune BART to generate more summary-worthy aggregations. These new approaches were fine-tuned for summarization using the same BART hyper-parameters reported in Lewis et al. (2020).

#### 4.3.1 BART+PretrainingAggregations

This approach tailors the pre-trained BART[5] towards aggregations before fine-tuning it on summarization. To do so, we further pre-train BART to reconstruct documents with masked aggregations. We expect this reconstruction knowledge to transfer to summarization.

The pre-training dataset comprises all the documents from the training set of CNN/DAILYMAIL. For each document, we mask aggregations (noun phrases filtered by TESA), plural noun phrases, and random spans until 30% of the tokens are masked. The BART checkpoint is further pre-trained during two epochs with batches of 64 samples to optimize the cross-entropy between the decoder's output and the original document.

#### 4.3.2 BART+AggregationChains

Recent works have shown that jointly learning to generate a sequence of summary-worthy entities followed by the summary can steer summaries towards those entities and also reduce hallucinations (Nan et al., 2021a; Narayan et al., 2021). We use a similar approach to encourage aggregations.

We fine-tune BART[5] to jointly generate the sequence of aggregations of the summary, fol-

[2] https://bit.ly/3fK3ZxU
[3] https://bit.ly/3tIHbXC
[4] https://bit.ly/3fFKXbV
[5] https://bit.ly/3AjKl5C

lowed by the summary. We built the dataset for fine-tuning BART from the training set of the CNN/DAILYMAIL corpus, discarding those samples whose summary has no aggregations. We fine-tuned the model during 20k steps with batches of 80 samples, as in (Lewis et al., 2020).

The target sequences for fine-tuning follow the format of Narayan et al. (2021), e.g., "[CHAIN] *rock climbers* | *friends* ||| *rivals* [SUMMARY] John and Mary are *rock climbers* and *friends*. They are also *rivals*.". During evaluation, the generated aggregation chain is removed.

#### 4.3.3 Gating BARTs

We found that improvements in aggregation production were typically accompanied with a deterioration in the summary content quality metrics. This difference is pronounced when evaluating on the whole CNN/DAILYMAIL test set (since our aggregation models are trained only on the subset which has aggregations). We aim to alleviate this inverse correlation by combining the best model at each evaluation level: BART+AggregationChains (aggregation level: the joint model just described) and BART$^2$ (summary level: a baseline summarization model).

We use BART+AggregationChains for summaries which contain aggregations, and BART$^2$ otherwise. We use a binary classifier to determine these cases. This classifier is a DEBERTA-LARGE[6] model, fine-tuned on the training set of the CNN/DAILYMAIL, to determine, given a document, if the reference summary has aggregations. We fine-tuned the classifier for six epochs using batches of 64 samples. This binary task can be done with an accuracy of 76% and an average F-score of 70.5. Table 11 of Appendix A shows the detailed results.

### 4.4 Oracles

We determine the upper bounds for our models using oracles of the above models. They were fine-tuned as described above for summarization. But during inference, essential information involving the aggregations of the reference summaries is disclosed as an input.

#### 4.4.1 BART+PerfectAggregationChains

This oracle observes how BART+AggregationChains would perform if it generated perfect aggregation chains. BART+AggregationChains first generates an aggregation chain, that is attended through decoder self-attention to condition the generation of the summary. Therefore, an upper bound on its performance is obtained by using as prefix for the decoder the chain with the oracle aggregations of the reference summary.

#### 4.4.2 Copy Sentence Oracles

BART, PEGASUS, and $T_5$ exhibit a strong behavior toward copying content from the source due to the high degree of extractivity in the CNN/DAILYMAIL corpus. We observe how these systems would behave if they could copy entire sentences with aggregations from the source into the summary.

We build three copy sentence oracles: Copy BART, Copy PEGASUS, and Copy $T_5$. Each sentence with aggregations in the reference summary is copied into the source document, and placed after the most similar sentence in the document in terms of averaged ROUGE scores. Then, each system is used to summarize the enriched document. Note that these oracles are much more informed than BART+PerfectAggregationChains since full sentences of the reference summary are disclosed.

## 5 Results

The results of the models and the oracles, at aggregation and summary levels, are presented in Tables 7 and 8 respectively. Note that COPY and NOVEL results are on a subset of the CNN/DAILYMAIL test set where aggregations are present in the summary. The full test set results are also reported separately.

All the models struggle at aggregation-level with the NOVEL test set, especially when they are evaluated using the most restrictive metric (Aggregation exact match). In this case, the models obtain approximately 10 times lower performance than with the COPY test set. Even with the COPY test set, where the models can copy the aggregations from the source to the summary, the results are almost always lower than 50 precision, recall, and $F_1$, which shows the difficulty of the task for current summarization approaches.

At the aggregation-level, BART+AggregationChains systematically outperforms all the other systems, showing that content planning with aggregation chains is an appropriate strategy to generate more summary-worthy aggregations. However, its performance at the summary-level is worse than that of BART, especially on the CNN/DAILYMAIL test set. This

[6] https://bit.ly/3sOh7cQ

| | COPY | | | NOVEL | | |
|---|---|---|---|---|---|---|
| | Aggregation Exact Match | Token Exact Match | Aggregation Relaxed Match | Aggregation Exact Match | Token Exact Match | Aggregation Relaxed Match |
| **Main Systems** | | | | | | |
| **BART** | **31.35** | **40.86** | **45.97** | **3.87** | **17.58** | **24.47** |
| **PEGASUS** | 29.32 | 38.25 | 42.97 | 3.71 | 16.40 | 22.15 |
| **$T_5$** | 28.12 | 37.36 | 43.39 | 3.79 | 16.20 | 22.67 |
| **Fine-tuned on Aggregations** | | | | | | |
| **BART+PretrainingAggregations** | 31.07 | 40.69 | 45.89 | 3.69 | 16.88 | 24.38 |
| **BART+AggregationChains** | **37.65** | **51.32** | **58.49** | **5.39** | **23.44** | **35.11** |
| **Gating BARTs** | 35.15 | 47.69 | 54.41 | 4.55 | 20.01 | 29.55 |
| **Oracles** | | | | | | |
| **Copy BART** | 44.11 | 51.41 | 55.71 | 22.17 | 34.18 | 39.25 |
| **Copy PEGASUS** | 43.18 | 50.36 | 54.30 | 18.64 | 30.72 | 34.94 |
| **Copy $T_5$** | 38.87 | 46.64 | 51.80 | 15.31 | 28.12 | 33.25 |
| **BART+PerfectAggregationChains** | **59.73** | **67.95** | **71.85** | **39.23** | **54.80** | **58.97** |

Table 7: Aggregation-level results of each model ($F_1$ scores) for each test set.

| | COPY | | | | NOVEL | | | | CNN/DailyMail test set | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R-1 | R-2 | R-L | BS | R-1 | R-2 | R-L | BS | R-1 | R-2 | R-L | BS |
| **Main systems** | | | | | | | | | | | | |
| **BART** | **46.11** | 22.89 | 43.03 | 31.73 | 42.95 | 18.70 | 39.60 | 28.19 | 44.06 | 21.07 | 41.00 | 30.53 |
| **PEGASUS** | 46.00 | **23.17** | **43.06** | **36.35** | **43.14** | **19.20** | **39.99** | **33.31** | **44.60** | **21.65** | **41.64** | **35.84** |
| **$T_5$** | 43.21 | 20.36 | 40.27 | 27.74 | 40.25 | 16.68 | 37.15 | 24.30 | 41.57 | 18.92 | 38.68 | 25.58 |
| **Fine-tuned on Aggregations** | | | | | | | | | | | | |
| **BART+PretrainingAggregations** | **46.04** | **22.73** | **43.06** | **35.21** | **43.10** | **18.85** | **39.99** | **32.18** | **44.05** | **20.96** | **41.13** | **34.25** |
| **BART+AggregationChains** | 45.14 | 21.94 | 42.12 | 34.71 | 41.96 | 18.00 | 38.86 | 31.25 | 42.02 | 19.05 | 38.93 | 31.95 |
| **Gating BARTs** | 45.08 | 21.84 | 42.06 | 34.20 | 42.19 | 18.22 | 39.03 | 30.56 | 43.32 | 20.37 | 40.30 | 31.57 |
| **Oracles** | | | | | | | | | | | | |
| **Copy BART** | 48.92 | 27.32 | 46.05 | 34.74 | **46.20** | 23.66 | 43.16 | 31.67 | 44.90 | 22.36 | 41.89 | 31.41 |
| **Copy PEGASUS** | **49.17** | **27.91** | **46.41** | **39.71** | 46.16 | **23.68** | **43.24** | **36.49** | **45.48** | **22.96** | **42.55** | **36.79** |
| **Copy $T_5$** | 45.54 | 23.86 | 42.82 | 29.97 | 42.57 | 19.83 | 39.71 | 26.49 | 42.22 | 19.88 | 39.40 | 26.22 |
| **BART+PerfectAggregationChains** | 46.22 | 23.97 | 43.42 | 33.35 | 43.41 | 19.98 | 40.53 | 30.76 | 43.87 | 20.97 | 41.02 | 33.76 |

Table 8: Summary-level results of each model ($F_1$ scores) for each test set, along with the CNN/DAILYMAIL test set. R stands for ROUGE and BS for BERTSCORE.

inverse correlation is alleviated by Gating BARTs, which trades off the aggregation-level and the summary-level performance better than BART and BART+AggregationChains. In this way, Gating BARTs represents an intermediate point between BART and BART+AggregationChains, that generally obtains better aggregation-level performance than BART and better summary-level performance than BART+AggregationChains. Regarding BART+PretrainingAggregations, its performance does not significantly differ from BART neither at aggregation-level nor at summary-level.

The upper bounds of the performance at aggregation and summary levels are posed by BART+PerfectAggregationChains and Copy PEGASUS oracles respectively. At the aggregation level, the performance of all the models is far below the upper bound posed by BART+PerfectAggregationChains, which suggests that there is great room for improvement of summarization systems. In addition, although the copy sentence oracles also have access to the reference aggregations, they obtain significantly worse results than the BART+PerfectAggregationChains. It suggests that summarization systems that have not been trained to consider aggregations properly will struggle in the aggregation-level evaluation. At the summary-level, the performance of the models is more similar to the oracles than in the aggregation-level evaluation, which suggests that there is smaller room for improvement here compared to aggregation-level.

Table 12 of Appendix A illustrates sample outputs from BART and BART+AggregationChains systems.

## 6 Conclusion and future work

We studied source-summary entity aggregation, a frequent phenomenon in the CNN/DAILYMAIL corpus. We analyzed the capabilities of state-of-the-art summarization systems to generate summary-worthy aggregations, and explored different ways of fine-tuning BART to generate more aggregations. Our results suggest that summarization models can improve greatly along these lines.

In future work, we would like to explore how to leverage knowledge about the entities to generate better aggregations. Another important direction is other types of semantic generalization, such as aggregations of sequences of events. Also, we plan

to explore source-summary entity aggregation on more abstractive summarization datasets such as XSUM and WIKIHOW, which could reflect better the aggregation phenomenon. Finally, we would like to investigate deeper questions about semantic aggregation in summarization, e.g., when is generality preferred over specificity in summaries?

## 7 Ethical considerations

The TESA dataset centers around specific topics found in the NEW YORK TIMES corpus during 2006-2007. They are skewed towards the male gender, and newsworthy entities involved in politics, business, etc. This selection limits the diversity of the aggregations used in our work. Even though models trained on the data learn semantic abstractions which aids generalization, we need further studies to explore how they differ in performance for different classes of entities.

Our models also share the same research issues as other abstractive systems and further work on reducing hallucinations, and factual inconsistencies will improve our approaches as well.

## Acknowledgements

We would like to thank Clément Jumel for his support in the initial steps of the project and for providing us the codebase and the TESA dataset.

This work has been partially supported by Ministerio de Ciencia e Innovación & FEDER funds under the project BEWORD PID2021-126061OB-C41.

## References

Riadh Belkebir and Ahmed Guessoum. 2016. Concept generalization and fusion for abstractive sentence generation. *Expert Systems with Applications*, 53:43–56.

Alicia Burga, Sergio Cajal, Joan Codina-Filbà, and Leo Wanner. 2016. Towards multiple antecedent coreference resolution in specialized discourse. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2052–2057, Portorož, Slovenia. European Language Resources Association (ELRA).

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Clément Jumel, Annie Louis, and Jackie Chi Kit Cheung. 2020. TESA: A Task in Entity Semantic Aggregation for abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8031–8050, Online. Association for Computational Linguistics.

Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. 2019. Abstractive text summarization based on deep learning and semantic content generalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5092, Florence, Italy. Association for Computational Linguistics.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for*

*Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O'Regan, Duygu Altinok, György Orosz, Søren Lind Kristiansen, Lj Miranda, Daniël De Kok, Roman, Explosion Bot, Leander Fiedler, Grégory Howard, Edward, Wannaphong Phatthiyaphaibun, Yohei Tamura, Sam Bozek, Murat, Ryn Daniels, Mark Amery, Björn Böing, Bram Vanroy, and Pradeep Kumar Tippa. 2022. spaCy: Industrial-strength Natural Language Processing in Python.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021a. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021b. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, and Ryan T. McDonald. 2021. Planning with entity chains for abstractive summarization. *CoRR*, abs/2104.07606.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sandhaus, Evan. 2008. *The New York Times Annotated Corpus*. In *Philadelphia: Linguistic Data Consortium. LDC2008T19. DVD.*

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3280–3291, Hong Kong, China. Association for Computational Linguistics.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Liu Ren, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Matthew Stone. 2000. On identifying sets. In *Proceedings of the First International Conference on Natural Language Generation - Volume 14*, INLG '00, page 116–123, USA. Association for Computational Linguistics.

Hardik Vala, Andrew Piper, and Derek Ruths. 2016. The more antecedents, the merrier: Resolving multi-antecedent anaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2287–2296, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2020. Free the plural: Unrestricted split-antecedent anaphora resolution. *CoRR*, abs/2011.00245.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Hao Zhou, Weidong Ren, Gongshen Liu, Bo Su, and Wei Lu. 2021. Entity-aware abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 351–362, Online. Association for Computational Linguistics.

# Appendix A Aggregations and system output

| Person | Location | Organization |
|---|---|---|
| (Joseph Lieberman, Ned Lamont) –> politicians | (Israel, Lebanon) –> middle eastern countries | (Airbus, Boeing) –> transportation companies |
| (Ehud Olmert, Mahmoud Abbas) –> politicians | (Iran, Iraq) –> neighboring countries | (London Stock, Nasdaq Stock) –> stock markets |
| (Charlie McDermott, Kris Kristofferson) –> actors | (Ethiopia, Somalia) –> african countries | (Microsoft, Google) –> technology companies |
| (Barry Diller, Frank Gehry) –> americans | (China, North Korea) –> asian countries | (Altimo, Telenor) –> companies |

Table 9: Several examples among the most frequent (entities, aggregation) pairs for each type in TESA.

| CC+PS | CC+IS | CC+NS |
|---|---|---|
| Internet giants signed up Tuesday to a "zero tolerance" approach to images of child sexual abuse as the British government announced a new, tougher strategy to find and block illegal content. Google, Yahoo, Microsoft, Twitter and Facebook were among the firms summoned to a meeting on the issue at 10 Downing Street, the prime minister's residence, by the UK government's Department for Culture, Media and Sport. | Internet giants signed up Tuesday to a "zero tolerance" approach to images of child sexual abuse as the British government announced a new, tougher strategy to find and block illegal content. Google, Yahoo, Microsoft, Twitter and Facebook were among the firms summoned to a meeting on the issue at 10 Downing Street, the prime minister's residence, by the UK government's Department for Culture, Media and Sport. | The launch of the lifeboat by William and Catherine and, at the same time, the launch of William and Catherine into this celebrity saturated world they are going to be living in. Despite the modest nature of the event, hundreds of people turned out to watch the royal couple conduct their first official duty together. |
| **CC+PL** | **IS+PS** | **IS+IS** |
| Last month, Inter was fined $65,500 by the Italian football authorities after its fans were found guilty of racially abusing former players Mario Balotelli and Sulley Muntari, who now play for rival AC Milan. | Some of the candidates have watched the video. Vazquez Mota, of the ruling National Action Party, said the video's message can't go unnoticed, while Institutional Revolutionary Party candidate Pena Nieto expressed that now is the time for change, as the video suggests. | As advertising losses and new reader habits afflicted newspapers nationwide, The Times began looking to shed The Globe and even threatened to close the paper in 2009 amid disputes with unions. |
| **IS+NS** | **CS+PS** | **CS+IS** |
| When asked what would happen if he rapped his anti-regime lyrics prior to Libyaś uprising, MC Swat said, "I would be shot to death like Tupac," referring to the American rapper killed in 1996.But here in Benghazi, the oppositionś de facto capital, thereś no sign of Gadhafiś loyalists anymore – or the fear that kept artists like MC Swat quiet for so long. | All three suspects are facing a charge of capital murder with the intent to sell a controlled substance, Lindley said. Trent Deundra Crump turned himself in to authorities of Alachua County Sheriff's Department in Gainesville, Florida, Lindley said. Duntae Harvey, 21, was arrested Monday and was being transferred Tuesday from Rankin County, where he has been held, university officials said. Mason Perry Jones, 21, of Jackson was arrested Monday in Memphis by members of the U.S. Marshal's Fugitive Task Force, Lindley said. | South Dakota, which has sent one inmate to death in three decades, has scheduled a lethal injection in October. Nebraska is the only state that does not use lethal injection, but its use of the electric chair was ruled unconstitutional in February. Texas and Mississippi are among the states that use 2 grams of sodium thiopental, the anesthetic used to render condemned inmates unconscious. |
| **CS+NS** | **SN+PE** | |
| Alexis was not the first mass killer to have an obsession with violent video games. Adam Lanza, who killed 26 children in an elementary school in Newtown, Connecticut, was also said to be a fan of first-person shooting games. Other killers have been found to be avid players. | Ivanovic's new team includes coach and hitting partner Nemanja Kontic – who represented Montenegro in the Davis Cup – fitness coach[es] Zlatko Novkovic and physio Branko Penic. They have all been part of her entourage since her split with British coach[es] Nigel Sears in July, following a second-round exit at Wimbledon. | |

Table 10: Alignments extracted by the heuristics in different examples from CNN/DAILYMAIL. The names are the acronyms of the patterns used to detect entities and aggregations.

| | P | R | $F_1$ | #samples |
|---|---|---|---|---|
| **Aggregation** | 51.31 | 66.58 | 57.96 | 2855 |
| **Not aggregation** | 87.75 | 79.11 | 83.29 | 8635 |
| **Macro-Avg** | 69.53 | 72.85 | 70.58 | 11490 |
| **Accuracy** | - | - | 76.00 | 11490 |

Table 11: Results of the classifier of Gating BARTs, per class and macro-averaged, on the test set of CNN/DAILYMAIL.

Table 12: Six examples of summarization using BART and BART+AggregationChains on the NOVEL and COPY test sets. The aggregations of the summaries are in green. They are also marked in the source if they are COPY aggregations. The entities of the document being aggregated (blue) are also shown, if they appear in the source, for visualization purposes.

Example 1 (NOVEL)

**Document**: (CNN)That sound you just heard was the crash of hearts breaking all over the world. Zayn Malik is leaving One Direction. "After five incredible years Zayn Malik has decided to leave One Direction," the band said on its Facebook page and tweeted out to its 22.9 million Twitter followers. "Niall, Harry, Liam and Louis will continue as a four-piece and look forward to the forthcoming concerts of their world tour and recording their fifth album, due to be released later this year." Rumors about such a move had started since Malik left the band's tour last week. At the time, a rep told Rolling Stone he had "been signed off with stress" after a scandal erupted following the publication of a photo showing Malik holding hands with someone other than his fiancee. Fans on Twitter immediately responded with teary Vine videos and the #AlwaysInOurHeartsZaynMalik hashtag. Even the Girl Scouts got in on the act with a sweet tribute to the singer. The band's Wikipedia page was also quickly updated with a sentence, "Zayn Malik was formerly a member." And Spotify said that global streams of One Direction songs were up 330% Wednesday in the hour after the news was announced – a "spike of sadness," as the music service called it. In the U.S. alone, streams of the band's music were up 769%. To mark the occasion, Spotify created a special playlist of 1D songs. Malik, 22, has been part of the very popular British boy band since it was formed (at the urging of Simon Cowell, according to some stories) in 2010 after members auditioned separately for the UK version of "The X Factor." Cowell became a mentor and signed them to his label. The group has put out four albums, and its hits include "Best Song Ever" and "Story of My Life." Their latest album, "Four," came out in November.

**Reference summary**: Malik had left the band's tour last week, citing "stress" One Direction is one of the most popular bands in the world .

**BART**: Zayn Malik is leaving One Direction. Rumors about such a move had started since Malik left the band's tour last week. Fans on Twitter immediately responded with teary Vine videos and the #AlwaysInOurHeartsZaynMalik hashtag. Spotify said that global streams of One Direction songs were up 330% Wednesday.

**BART+AggregationChains**: Zayn Malik is leaving One Direction . He has been part of the british boy band since it was formed . Spotify says global streams of One Direction songs were up 330% Wednesday in the hour after the news was announced .

Example 2 (NOVEL)

**Document**: (CNN)Comic books of the past few years have seen a lot of changes (a female Thor, anyone?) but not quite so many at one time. Three major characters – Superman, Wonder Woman (both of DC Comics, a Time Warner company, like CNN) and Archie Andrews – came out with new looks (and costumes in two cases) Thursday. Superman and Wonder Woman are no stranger to change over time, but these are pretty different from what we're used to. Wonder Woman is back to wearing pants (similar to her style circa 2010, not to mention a brief time in the 1970s) along with something of a turtleneck and body armor. Superman looks the most casual that we've ever seen him, simply in a t-shirt and jeans, and decidedly shorter hair. He also looks like someone you might not want to come across in a dark alley. As for Archie (who never really died, by the way), he's getting ready for his promised TV show by debuting a modern look – one that makes it much easier to see why Betty and Veronica have been fighting for his affections all of these years. It's been a whirlwind time for comic book fans, and there will be a lot to get used to.

**Reference summary**: Superman, Wonder Woman and Archie all debuted new looks Thursday . Three of the most well-known comic book characters of all time look radically different .

**BART**: Superman, Wonder Woman and Archie Andrews have new looks . It's been a whirlwind time for comic book fans, and there will be a lot to get used to . For more comic book news, visit CNN.com/comic-book .

**BART+AggregationChains**: Three major characters – Superman, Wonder Woman and Archie Andrews – came out with new looks (and costumes in two cases) Superman looks the most casual that we've ever seen him, simply in a t-shirt and jeans. Archie (who never really died, by the way) is getting ready for his promised TV show.

Example 3 (NOVEL)

**Document**: (CNN)Italian authorities have arrested 42 people suspected of belonging to a mafia network in southern Italy, some of whom are accused of extorting tens of thousands of euros from businesses monthly, police said Tuesday. Among the evidence: A piece of paper, hidden in a bicycle handlebar, that listed businesses that the crime syndicate was extorting, according to the Carabinieri, Italy's military police. The Carabinieri said they made the arrests while raiding homes in 11 provinces. The crime network is based in Casal di Principe, a village roughly 20 kilometers (12 miles) northwest of Naples, police said. The mafia network extorted shop owners to the tune of 200,000 euros a month and extorted operators of slot machines and online gambling a further 100,000 euros monthly, the Carabinieri said. The arrests stem from an investigation that started in October 2012. Investigators also found notebooks with the group's organizational chart and names of jailed associates who still were on the mafia payroll, according to the Carabinieri. The note in the bicycle handlebar was among other "pizzini," or little pieces of paper allegedly used by mafia members to exchange information, that investigators discovered in the probe, the military police said. In Villa Literno, close to Casal di Principe, investigators discovered a bunker that allegedly was used to hide fugitive affiliates of the mafia, the Carabinieri said. An anti-mafia prosecutor in Naples filed charges against the 42, including mafia association, extortion and illegal possession of weapons.

**Reference summary**: Police say a note hidden in a bike's handlebar listed extorted businesses . The crime network is based in Casal di Principe, a village near Naples, police say . Arrests stemmed from investigation that began in 2012 .

**BART**: Italian authorities arrest 42 people suspected of belonging to a mafia network in southern Italy. A piece of paper, hidden in a bicycle handlebar, listed businesses that the crime syndicate was extorting. The crime network is based in Casal di Principe, a village roughly 12 miles northwest of Naples.

**BART+AggregationChains**: A piece of paper, hidden in a bicycle handlebar, listed businesses that the crime syndicate was extorting . The crime network is based in Casal di Principe, a village roughly 20 kilometers (12 miles) northwest of Naples .

Example 4 (COPY)

**Document**: A game of cat and mouse has been captured in a series of striking images as the pair battle it out on a shed rooftop like a real life version of much-loved cartoon duo Tom and Jerry. It is an age-old rivalry that rarely ends well for one of its parties and so it proved in this remarkable set of photos. The snaps of a cat playing with a mouse on a roof in Shepton Mallet, Somerset, illustrate the perils the tiny rodents face in the town. Ironically the pet cat's name is Mouse. Unfortunately for this mouse that's where all similarities between the moggy and its namesakes end. The pictures were taken by the cat's owner Jason Bryant who confirmed the inevitable outcome of the encounter. 'My cat is a very good mouser,' he said. 'She's done it before. She often brings them into the house at 3am and then I can rescue them. But she took this poor little thing on to the shed roof where I couldn't reach it.' Ready to pounce: The mouse flies through the air as its feline nemesis looks to bring it back down to earth with a bang . Eeek: A mouse tries to escape the clutches of a pet cat named Mouse as the pair engage in the age-old game of cat and mouse . In flight: The mouse, unable to evade its moggy hunter, is thrown into the air on the roof of a shed in Shepton Mallet in Somerset . Somersaulting away: The snaps illustrate the perils that tiny rodents face in the town with cats like Mouse around . Cartoon characters: The pair battle it out on a rooftop like a real life version of much-loved cartoon duo Tom and Jerry . Let's play: The pictures were taken by the cat's owner Jason Bryant who confirmed the inevitable outcome of the encounter . Reaching out: The mouse is flung high in the sky as it finally admits defeat in the game of cat and mouse .

**Reference summary**: Age-old game of cat and mouse is brought to life in these quirky pictures taken in Shepton Mallet in Somerset . The pair are seen battling it out on the roof of a shed in a real life take on an episode of Tom and Jerry . Ironically, the cat's name is Mouse. The pictures show the dangers small rodents have to be aware of in the area .

**BART**: The snaps of a cat playing with a mouse on a roof in Shepton Mallet, Somerset, illustrate the perils the tiny rodents face in the town. Ironically the pet cat's name is Mouse. The pictures were taken by the cat's owner Jason Bryant who confirmed the inevitable outcome of the encounter.

**BART+AggregationChains**: The snaps of a cat playing with a mouse on a roof in Shepton Mallet, Somerset, illustrate the perils the tiny rodents face in the town . Unfortunately for this mouse that's where all similarities between the moggy and its namesakes end . The pictures were taken by the cat's owner who confirmed the inevitable outcome of the encounter .

Example 5 (COPY)

**Document**: (Billboard)Fresh off his scorching performance at Coachella Saturday night (and days before his next one on the festival's second weekend), rocker Jack White announced he'll take a hiatus from touring. White will wrap his touring efforts in support of "Lazeretto" with a brief, first-ever acoustic tour that will hit "the only five states left in the U.S. that he has yet to play," according to White's website. Rounding out the acoustic quartet on tour will be Fats Kaplin, Lillie Mae Rische and Dominic Davis. The shows will be unannounced until day-of-show, with tickets priced at $3 and limited to one ticket per person, to be purchased only at the venue on a first-come, first-served basis. Billboard: Jack White on Not Being a 'Sound-Bite Artist,' Living in the Wrong Era and Why Vinyl Records Are 'Hypnotic' The purposely vague announcement surely has fans (and journalists) scouring the Internet for White's touring history. Unclear is whether White includes his work with The White Stripes, The Raconteurs and Dead Weather in his touring history, or just his solo road work. Presumably, he's including all of his touring, with all bands, as Billboard could find only 29 states in which he has performed as Jack White. Tour dates with White Stripes add another 12 states. That leaves nine states for which we could not find a show for White: Hawaii (where a show is scheduled for tomorrow, April 15), Arkansas, Idaho, Utah, Wyoming, Vermont, Iowa, and North and South Dakota. Billboard: Jack White Plays The Hits, Declares 'Music Is Sacred' at Coachella . Through the process of elimination (surely he has played Boise, Little Rock, and Salt Lake?), our guess as to which five states White will play on the brief acoustic run: South and North Dakota, Wyoming, Vermont and ... Puerto Rico? If that's the case, this tour is in for some long jumps, with Puerto Rico to Vermont being a potential beast. (Though shipping acoustic instruments and ribbon mics will be a lot less taxing than a full electrified stage setup.) ©2015 Billboard. All Rights Reserved.

**Reference summary**: Jack White taking a hiatus from touring after brief acoustic jaunt . He'll play five states he has yet to get to, charge just $3 . Places and times of shows are currently a mystery .

**BART**: Jack White announced he'll take a hiatus from touring. White will wrap his touring efforts in support of "Lazeretto" with a brief, first-ever acoustic tour. The shows will be unannounced until day-of-show, with tickets priced at $3.

**BART+AggregationChains**: White will wrap his touring efforts in support of "Lazeretto" with a brief, first-ever acoustic tour . He will hit "the only 5 states left in the U.S. that he has yet to play," according to White's website . The shows will be unannounced until day-of-show, with tickets priced at $3 .

Example 6 (COPY)

**Document**: Boss Nigel Pearson has urged Leicester to keep their cool and ignore their relegation rivals. The Foxes host Swansea on Saturday just three points from safety in the Barclays Premier League after back-to-back wins. Last week's 3-2 win at West Brom handed them a survival lifeline, although they remain bottom of the table. Jamie Vardy scored an injury-time winner against West Bromwich Albion on Saturday to improve his side's slim chance of Premier League survival . Vardy celebrates in front of the travelling away fans after hitting the winner against West Brom . But after their mini-revival, Pearson wants his side to remain focused on their own jobs. 'I'm very wary of people flipping the emphasis,' he said. 'Our future is in our own hands and if we go into the last game with that we have given ourselves a realistic chance. 'We need to make sure our own run-in is what we want it to be. Leicester manager Nigel Pearson has urged his players to focus on their own job and not worry about their relegation rivals during run-in . 'It's helpful that other teams are being dragged into it. People are continually adjusting the form table and their predictions. 'We can't get involved in anything apart from being involved in our own games that may change people's perceptions. 'It's great when results go your way but they only help if you've done your own job. 'It's important for us to be in touch. One of the potential problems we have to deal with is people are now overly optimistic.'

**Reference summary**: Leicester have won back-to-back league games to boost survival hopes . Nigel Pearson has urged his players to focus on their own run-in . Leicester now just three points from safety heading into final six games .

**BART**: Nigel Pearson has urged Leicester to ignore their relegation rivals. The Foxes host Swansea on Saturday just three points from safety. Last week's 3-2 win at West Brom handed them a survival lifeline. But Pearson wants his side to remain focused on their own jobs.

**BART+AggregationChains**: Nigel Pearson has urged Leicester to keep their cool . The Foxes host Swansea on Saturday just three points from safety in the Barclays Premier League after back-to-back wins . But after their mini-revival, Pearson wants his side to remain focused on their own jobs .

## Appendix B Notes on the heuristics

Through a preliminary evaluation, we observed that **Coordinating Conjunction+Preceding a list** (**CC+PL**) and **Shared noun+Preceding Entities** (**SN+PE**) stand out in terms of precision. Especially, **CC+PL** has almost 100% precision and appears in 3.87% of the documents in the CNN/DAILYMAIL corpus (13,959 alignments). **SN+PE** is less frequent than **CC+PL** (1.27% of documents and 4125 alignments). The other heuristics ranges from 20% to 80% of precision, and some of them such as **In Sentence+In span** seems to have a high recall (25.74% of documents and 132,922 alignments).