TESA: A Task in Entity Semantic Aggregation for Abstractive Automatic Summarization

Clément Jumel
McGill School of Computer Science,
McGill University, Montreal

August, 2020
A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master in Computer Science (Thesis)

Contents

Abstract

During the past years, the increasing amount of textual data available, coupled with the development of new machine learning technologies, has led to significant progress in many fields of natural language processing, such as automatic summarization, which now leads to summaries always closer to human standards. However, in general, human-written texts contain frequent generalizations and semantic aggregation of content. For example, in a document, they may refer to a pair of named entities such as 'London' and 'Paris' with different expressions: "the major cities", "the capital cities" or "two European cities". Yet, automatic text generation, especially, abstractive automatic summarization systems have so far focused heavily on paraphrasing and simplifying the source content, to the exclusion of such semantic abstraction capabilities.

In this thesis, we present a new dataset and task, named TESA, aimed at the semantic aggregation of *entities*. TESA contains 5.3K crowd-sourced entity aggregations of PERSON, ORGANIZATION, and LOCATION named entities. The aggregations are document-appropriate, meaning that they are produced by annotators to match the situational context of a given news article from the New York Times. We then build baseline models for generating aggregations given a tuple of entities and document context. We fine-tune an encoder-decoder language model on TESA and compare it with simpler classification methods based on linguistically informed features. Our quantitative and qualitative evaluations show reasonable performance in making a choice from a given list of expressions, but free-form expressions are understandably harder to generate and evaluate.

Résumé

Au cours des dernières années, la quantité croissante de données textuelles disponibles, associée au développement de nouvelles technologies d'apprentissage machine, a conduit à des progrès significatifs dans de nombreux domaines du traitement automatique des langues, tels que le résumé automatique de texte, qui permet désormais de générer des résumés de plus en plus proches des standards humains. Cependant, les textes d'origine humaine contiennent en général des généralisations et des aggrégations sémantiques de leur contenu. Par exemple, dans un document, un texte peut faire référence à une paire d'entités nommées comme 'Londres' et 'Paris' avec diverses expressions telles que "les villes majeures", "les capitales" ou "deux villes européennes". Cependant, la génération automatique de texte, et plus particulièrement les systèmes de résumé automatique abstractifs, se sont concentrés largement sur la paraphrase et la simplification du document source, en excluant de telles capacités d'abstraction sémantiques.

Dans cette thèse, nous présentons une nouvelle base de données et une tâche, nommées TESA, visant à l'aggrégation sémantique d'*entités*. TESA contient 5.3K aggrégations d'entités, crowdsourcées à partir d'entités appartenant aux catégories de PERSONNE, ORGANISATION ou EMPLACEMENT. Les aggrégations sont spécifiques à chaque document, c'est-à-dire qu'elles sont produites par des annotateurs pour correspondre au contexte d'un article de presse donné du New York Times. Nous avons ensuite développé des modèles simples d'apprentissage automatique pour la génération d'aggrégations, étant donné un ensemble d'entités et un document de contexte. Nous avons entraîné sur TESA un modèle encodeur-décodeur de langage pré-entraîné, et nous l'avons comparé à des méthodes de classification plus simples s'appuyant sur des idées intuitives de langue. Nos évaluations quantitatives et qualitatives démontrent des performances raisonnables pour faire un choix depuis une liste d'expressions, cependant la génération et l'évaluation d'expressions libres est plus difficile, comme attendu.

Acknowledgements

This thesis is the fruit of a great collaboration with my two amazing supervisors, Annie Louis and Prof. Jackie C. K. Cheung, who I thank for their kindness, patience and diligence, and, last but not least, for their bright ideas in Natural Language Processing, which guided me through the whole project.

I also want to thank the Mila - Institut québécois d'intelligence artificielle, for providing me with a great work environment, as well as many classes and reading groups to help me expand my scientific horizons. I would also like to mention my friends from the lab, as they gave me their support during the entire project, and our Natural Language Processing group, lead by Prof. Jackie C. K. Cheung, which helped me during my work and shared great ideas and projects.

Finally, I wish to mention my two Guinea pig annotators, who did a very conscientious work in giving me feedback and helping me design the annotation process which lead to the creation of the TESA dataset through crowd-sourcing.

1 Introduction

The new era of deep learning, with the development of large and complex neural architectures, has led to an upheaval in many different domains of machine learning. For instance, natural language processing (NLP) has leveraged these new technologies, such as Transformers (Vaswani et al., 2017), as well as the large amount of textual data available in new datasets, in order to develop many new successful methods. One of the many subfields of NLP which benefited from this trend is automatic summarization, which aims at creating a short summary from an input text.

Nowadays, automatic summarization becomes more useful everyday, as more and more news articles are written for instance, through an increasing number of different media. Automatic summarization can thus become a mean to escape this overwhelming amount of information available by synthesizing the information found on the Internet and in books, and will probably find a lot of new applications in the near future.

Automatic summarization is performed typically with one of two main approaches. On the one hand, extractive summarization consists of solely copying and concatenating relevant and informative excerpts of the input text, which can be very efficient but remains limited, as it cannot perform actual abstraction or high level rewriting. On the other hand, abstractive summarization relies on the free generation of a summary, which is generally a much harder task to perform. Yet, abstractive summarization does not suffer from the inherent limitations of extractive summarization, and is therefore a more potent and human-like approach. Still, unlike some other subfields of machine learning (e.g. image recognition), abstractive summarization is still far from human-like performances and there is a lot of room for progress on many specific points, such as the abstraction capacity.

1.1 TESA

Abstractly speaking, abstraction can be defined as the process of deriving general concepts from specific instances. In automatic summarization, however, “abstractive” summarization often means any type of rewriting of words in some source document into an output summary. Con-

cretely, recent summarization datasets including XSum (Narayan et al., 2018) and NEWSROOM (Grusky et al., 2018) quantify the degree of abstractiveness of a summary in terms of its novel N-grams.

While such a surface-level definition of abstractiveness is certainly useful and convenient, it is nevertheless only a proxy for abstraction in the broader sense which concerns semantic generalization. We argue that it is important to also focus explicitly on semantic abstraction, as this capability is required for more difficult types of summarization which are out of reach of current methods. For example, generating a plot summary of a novel might require describing sequences of events using one sentence. Writing a survey of a scientific field would require categorizing papers and ideas, and being able to refer to them as a whole. Outside of domain-specific settings such as opinion summarization (Ganesan et al., 2010; Gerani et al., 2014, *inter alia*), and tasks such as sentence fusion (Barzilay and McKeown, 2005), there has been little work focusing on semantic generalization and abstraction.

In this thesis, we start to tackle this issue by focusing on the specific task of semantic aggregation of entities, that is, how to refer to a tuple of named entities using a noun phrase instead of enumerating them. To do so, one needs some entities to aggregate together and a context; in Table 1, we present a few examples of such aggregation.

Input Entities	François Bayrou, Nicolas Sarkozy, Ségolène Royal
Document Context	François Bayrou, Nicolas Sarkozy, and Ségolène Royal are the main contenders in the French presidential elections.
Possible Aggregations	• the French politicians • the French presidential candidates • the politicians

Table 1: An example of semantic entity aggregation. The input consists of a tuple of named entities, a situational (document) context, and background information about the entities (not shown here). The expected output is an aggregation of the tuple of entities.

We define a task to evaluate summarization models on semantic entity aggregation, which we call **TESA** (A **T**ask in **E**ntity **S**emantic **A**ggregation). In TESA, a system is presented with a

list of named entities in an original textual context, and it must produce a *non-enumerating* noun phrase which refers to the designated entities. Solving this task requires finding a semantic link between all the entities in the list (e.g., London and Paris are cities of considerable sizes), then using this information to generate a noun phrase (e.g., "the major cities").

We introduce an accompanying dataset of entities in context drawn from the New York Times (NYT) corpus (Sandhaus, 2008), and their aggregations which were written by crowd workers. Our dataset contains 5.3K aggregation expressions. Each example, contains a tuple of PERSON, ORGANIZATION or LOCATION named entities, a paragraph context from an NYT article discussing the entities, and background information about entities in the form of summary snippets from Wikipedia. We also introduce the first models for the TESA task which are based on an encoder-decoder system pretrained for abstractive summarization, BART (Lewis et al., 2019). We present two ways of fine-tuning BART to TESA, either in a discriminative or in a generative fashion, and compare them against simpler statistical and frequency-based methods.

The simple classifier achieves decent results on TESA. It is however outperformed by a wide margin by BART, when fine-tuned on our task in a discriminative manner. When fine-tuned as a generative model, BART yields similar performance as the simple classifier. Yet, the generative model is able to freely generate entity aggregations with diversity and quality, which represents a significant advantage compared to its discriminative counterpart for real world applications, despite some factual inconsistencies.

1.2 Thesis outline

Chapter 2 provides the scientific background needed for this thesis. We first explore some background on automatic summarization; that is, current methods, datasets and metrics, and details on the state of the art. We then focus on entity aggregation, and its links with several NLP tasks.

Chapter 3 introduces the novel TESA dataset. We explain in details our ideas behind entity aggregation and introduce what we call an *aggregatable instance*. We then detail how we built the dataset through data extraction first, and then crowd-sourcing.

Chapter 4 introduces the TESA task, which relies on the dataset we built. We describe our design choices to build a modeling task upon our dataset, introduce a few metrics and present some examples and statistics.

Chapter 5 presents the models we used to try and solve the TESA modeling task. We first present very simple baselines, useful to reckon the difficulty of the task, and then how we fine-tuned an encoder-decoder-based large pretrained model to apply state-of-the-art method to our task.

Chapter 6 exhibits the results of our models on the task. We display the raw scores as well as some qualitative examples and try to understand the limitations of our models.

Chapter 7 draws the conclusions from this thesis and explore a few ideas for future work.

1.3 Statement of contribution

This thesis is based on original ideas from my supervisors, Annie Louis and Jackie C. K. Cheung, which I tried to complete with my insights and my investment in this project. Besides, the experiments for creating the dataset and running the different models on the task have been carried by myself.

The TESA dataset presented in Section 3 and the corresponding task detailed in Section 4 are entirely original contributions which will be publicly available. The experiments performed on the task detailed in Section 5, and their results, detailed in Section 6, are also original contributions, despite relying on an existing model, BART (Lewis et al., 2019).

All the experiments and results of this thesis also appear in a paper submitted to the *2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2020), under the same title.

2 Background

This thesis is written in the context of the study of abstractive summarizers in automatic summarization. Therefore, we will start by giving a brief background of this field. Then, we will focus on the entity aggregation and introduce a few related tasks.

2.1 Automatic Summarization

Automatic summarization is a subfield of natural language processing (NLP) and natural language understanding (NLU), which are themselves subfields of artificial intelligence and machine learning. Automatic summarization typically involves a document as input (e.g., a press article), and the desired output is a short summary, which ideally is relevant and as much informative as possible.

Automatic summarizers typically follow one of two possible approaches: extractive summarization or abstractive summarization. Extractive summarization consists of the concatenation of the most relevant excerpts of the input document. Thus, it does not involve any text generation. It is an efficient method, but it has inherent limitations such as its inability to perform any kind of rewriting, including abstraction, therefore, this kind of method will not be of interest in our work.

On the other hand, abstractive summarization consists of fully generating the output summary. It is thus a much more difficult task, as it involves at the same time understanding the input document, synthesizing it, and generating the output in natural language. However, theoretically, it does not suffer from any form of intrinsic limitation.

2.1.1 Datasets

To tackle such a difficult task, many methods, for both extractive and abstractive summarization, involve machine learning and deep learning. Therefore, the need for large summarization datasets has arisen, leading to new datasets, such as the CNN/DailyMail dataset (Nallapati et al., 2016) (hundreds of thousands pairs of press article and multiple-sentence summary), or Gigaword (Rush et al., 2015) (millions of pairs of single-sentence input and headline summary).

Some datasets have also been developed with specific goals, such as the XSum (Extreme Summarization) dataset (Narayan et al., 2018). This dataset is designed to have extremely concise summaries, which is supposed to favor abstractive methods, inasmuch as the ideas of the summaries are supposedly very high-level and therefore cannot be solely extracted from the input document.

2.1.2 Metrics

To compare the performances of different methods on summarization datasets, many metrics have been developed to compare the output of a model and the gold standard summary. Given the nature of the summarization task, since two very different summaries can be both good summaries, human evaluation can often be preferred, even if it also has drawbacks. Especially, it can be expensive, hence the need of automatic metrics.

One of the most salient group of metrics is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004). ROUGE metrics typically measures the similarity of the output document and the gold standard through the overlap of N-grams (i.e. of any contiguous sequence of N tokens). It is quite convenient as it is an automatic metric; however, unlike (expensive) human evaluation, ROUGE metrics will fail to detect that two texts are close if they share the same semantic content but are made of different words. Common metrics of ROUGE are ROUGE-1 (which measures the overlap of unigrams), ROUGE-2 (which measures the overlap of bigrams) or ROUGE-L (which measures the longest common subsequence).

2.1.3 State of the art

Abstractive summarizers have gained prominence with the popularization of recurrent neural networks (RNNs) (Sutskever et al., 2014; Nallapati et al., 2016), and more recently Transformers (Vaswani et al., 2017), a neural architecture based on self-attention mechanisms. Indeed, Transformers have formed the basis of many new models, such as the famous BERT (Devlin et al., 2019), a large pretrained Transformer-based model, which have achieved the state of the art in a wide variety of NLP and NLU tasks. In addition, BERT also illustrated the recent success of a learning paradigm: pre-training a model on huge amount of textual data (e.g. BERT

was pre-trained on BooksCorpus (Zhu et al., 2015) and English Wikipedia) and fine-tuning the model on the task at hand. BERT's success is in part due to its pre-training objectives: the masked language model (MLM) and the next sentence prediction (NSP) objectives.

In BERT's wake, several abstractive models have achieved state-of-the-art performances on benchmark summarization datasets in terms of ROUGE, including ProphetNet (Yan et al., 2020), PEGASUS (Zhang et al., 2019) and BART (Lewis et al., 2019). BART is very similar to BERT as it follows the same pre-training and fine-tuning paradigm with a similar model architecture, but it benefits from a wider range of pre-training objectives, which makes it more appropriate for generation tasks such as automatic summarization.

Recent work has also focused on specific issues such as preventing inappropriate repetition (Kryściński et al., 2018), word-level rewriting, and evaluating factual consistency (Kryściński et al., 2019; Maynez et al., 2020).

2.1.4 Abstraction in automatic summarizers

Abstraction is critical for certain domains and applications, but has not been thoroughly explored in many. For example, in scientific article summarization, the particular structure and length of scientific articles make extractive techniques much easier to apply (Agarwal et al., 2011). Therefore, abstractive summarizers (Lloret et al., 2013) remain a minority. In opinion summarization, there have been abstractive systems that leverage cues specific to this task, such as redundancy in opinions (Ganesan et al., 2010) and specific discourse structures (Gerani et al., 2014). As abstractive systems have become strong in terms of generation capabilities, the time is apt to examine issues in semantic abstraction that could be useful in many summarization domains and tasks. Our work is a step in this direction.

2.2 Entity aggregation

In this thesis, we focus on a particular form of abstraction: entity aggregation. To the best of our knowledge, no previous work has directly addressed entity aggregation. Entity aggregation is a common phenomenon in texts, and can be similar to several different tasks of NLP and NLU, some of which we will describe here. However, even though entity aggregation is similar

to those tasks, we will see why we cannot use any of them directly in order to perform entity aggregation and thus, why needed to perform crowd-sourcing to create a quality and unbiased dataset.

2.2.1 Referring Expression Generation

Our proposed entity aggregation task is related to referring expression generation (REG). In REG, a system is given some targets (for example: "François Bayrou", "Nicolas Sarkozy" and "Ségolène Royale") and the system has to create a referring expression (noun phrase, pronoun, etc.) which identifies the targets.

REG is concerned with determining the form and content that entity references should take during generation (Krahmer and van Deemter, 2012; Castro Ferreira et al., 2018; Cao and Cheung, 2019). It emphasizes finding the right distinguishing characteristics of the intended referent or referents. Our work can be seen as a specific REG task that focuses on semantically abstracting multiple named entities.

2.2.2 Coreference resolution

A parallel can also be drawn between entity aggregation and coreference resolution. Coreference resolution is the task of linking any referring expression to the corresponding entity, or entities in the case of multi-antecedent resolution (Burga et al., 2016; Vala et al., 2016). For instance, in coreference resolution, when presented a text and a referring expression, such as "the politicians", a system has to find other referring expressions from the text corresponding to some of the entities designated by the input referring expression (for instance "François Bayrou", or "Ségolène Royale", if they are designated by "the politicians").

In that aspect, coreference resolution can be seen as an inverse problem to ours, inasmuch as, with entity aggregation, we want to generate a referring expression given several entities as antecedents.

2.2.3 Knowledge graphs

Finally, links can also be made between entity aggregation and knowledge graphs (KGs). Indeed, KGs, such as Yago (Suchanek et al., 2007) or DBpedia (Auer et al., 2007), are built on data extracted from a knowledge base, such as Wikipedia, and are a rich source of information. Especially, KGs can be used in order to retrieve hypernymns of entities (e.g. politician is an hypernym of Georges W. Bush). Given several entities, many common hypernym can be used to create an entity aggregation. Therefore, in theory, for a single set of entities, KGs could be used to generate many entity aggregations. However, using KGs, we cannot have information on the relevance of the information used to create the aggregation. For instance, Georges W. Bush is a man, an American, a former politician, a businessman, etc. KGs cannot help us in choosing the right information to use in order to create a relevant and natural entity aggregation.

3 The TESA dataset

In order to create our entity aggregation dataset, we first establish the ingredients needed for an entity aggregation, what we call in the following an **aggregatable instance**. Then, we describe how we use the New York Times (NYT) Annotated Corpus (Sandhaus, 2008) to extract the aggregatable instances' associated data. The NYT corpus contains high-quality metadata listing the salient named entities mentioned in each article. For instance, we form our tuples from entities tagged in the metadata for the same article. Finally, we detail how we used the aggregatable instances to collect entity aggregations during our crowd-sourcing experiments.

3.1 An aggregatable instance

An aggregatable instance, as we call it in this thesis, contains all the ingredients that we assume are needed for someone to write an entity aggregation. We will use the aggregatable instances we create in order to collect the entity aggregation during the crowd-sourcing experiments.

The starting point of an aggregatable instance is the tuple of named entities which should be aggregated and the type of its entities (e.g., PERSON). As we aim for contextual entity aggregations, an aggregatable instance also contains a document context; i.e., a passage from a document in which all the entities are mentioned. To provide additional background knowledge, we also include introductory summaries for the entities taken from Wikipedia. An example of aggregatable instance, as presented to the annotators during the crowd-sourcing experiments, is in Figure 1. For several aggregatable instances with the full information, see Table 2. Note that in the following, the examples we will display will evolve around the same three aggregatable instances, for simplicity and clarity purposes.

Figure 1: Example of aggregatable instance, displayed with the layout used during the crowdsourcing experiments. The mentions of the entities in the New York Times article are colored, and the name of the corresponding entity is available when an annotator clicks on a mention. The title of the Wikipedia information is an hyperlink to the corresponding Wikipedia page. The instructions of the annotation task are accessible through the corresponding button.

Aggregatable instance

Input entities Francois Bayrou, Nicolas Sarkozy and Segolene Royal
Entity type person
Background information

Francois Bayrou: François Bayrou is a French centrist politician and the president of the Democratic Movement , who was a candidate in the 2002, 2007 and 2012 French presidential elections.

Nicolas Sarkozy: Nicolas Paul Stéphane Sarközy de Nagy-Bocsa ; born 28 January 1955) is a retired French politician who served as President of France and ex officio Co-Prince of Andorra from 16 May 2007 until 15 May 2012.

Segolene Royal: Ségolène Royal ; born 22 September 1953), is a French politician and former Socialist Party candidate for President of France.

Context Street Violence by Paris Youths Intrudes Again Into French Politics: **The Socialist candidate, Ségolène Royal**, who is running second in the opinion polls, said the incident showed that **Mr. Sarkozy** had failed as interior minister." In five years with a right-wing government that has made crime its main campaign issue, you can see that it is a failure all the way," she said on Canal+ television. **François Bayrou, a centrist presidential candidate**, also took aim at Mr. **Sarkozy**, saying," It is very important to end this climate of perpetual confrontation between police and some citizens."

Input entities Chicago and London
Entity type location
Background information

Chicago: Chicago , locally also), officially the City of Chicago, is the most populous city in the U.S. state of Illinois and the third most populous city in the United States. With an estimated population of 2,705,994 , it is also the most populous city in the Midwestern United States. [...]

London: London is the capital and largest city of England and the United Kingdom. Standing on the River Thames in the south-east of England, at the head of its 50-mile estuary leading to the North Sea, London has been a major settlement for two millennia. [...]

Context Virtually Cool: The author of the hour was Chris Anderson, who after the drinks entertained the crowd with a simulcast PowerPoint lecture on the topic of his new best seller," The Long Tail," which describes how the chokehold of mass culture is being loosened by the new Internet-enabled economics of niche culture and niche commerce. The party was sponsored in part by a small SoHo-based new-media company called Flavorpill, which produces free e-mail magazines and weekly event guides for New York, Los Angeles, San Francisco, **Chicago** and **London**.

Input entities Microsoft Corp. and Sony Corp
Entity type organization
Background information

Microsoft Corp.: Microsoft Corporation is an American multinational technology company with headquarters in Redmond, Washington. It develops, manufactures, licenses, supports, and sells computer software, consumer electronics, personal computers, and related services. Its best known software products are the Microsoft Windows line of operating systems, the Microsoft Office suite, and the Internet Explorer and Edge web browsers. [...]

Sony Corp.: Sony Corporation is a Japanese multinational conglomerate corporation headquartered in Kōnan, Minato, Tokyo. Its diversified business includes consumer and professional electronics, gaming, entertainment and financial services.

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.

Table 2: Examples of aggregatable instances. An aggregatable instance contains the names of the input entities, their type, the background information extracted from Wikipedia and the New York Times article's context (title is underlined and the entities' mentions are in bold). For displaying purposes, these examples have been shortened.

3.2 Data extraction

While we could have gathered naturally occurring entity aggregations, work on multi-antecedent coreference resolution is still nascent, and our initial attempts to define heuristic methods to extract entity aggregations were very noisy. We instead used crowd-sourcing to gather human-generated aggregations from the aggregatable instances we extracted from the NYT corpus.

We used the 2006 and 2007 portions of the NYT corpus. We started with the editorial metadata which tags salient named entities in each article. These are entities we believe are likely to be included in a summary. We filtered the entity tuples to remove those that are unlikely to be naturally aggregatable using the following two constraints. First, the entities should have the same type (PERSON, LOCATION, or ORGANIZATION in this corpus). Second, the entities should be mentioned close together, within a span of consecutive sentences of the same length as the size of the tuple of entities (e.g., three consecutive sentences for three entities). We also selected those entity tuples that are mentioned together in the abstract of an article.

To extract the document context, we extracted both the title of the article and the span of sentences which mentions the entities. If the same entity tuple is mentioned in different qualifying sentence spans in the same article, they would be extracted as different aggregatable instances.

As for the background information, we extracted an excerpt of each entity's Wikipedia article, using the first paragraph of the article if it exists, up to 600 tokens. We used the entity name to identify its Wikipedia page[1], and, in case of ambiguous or incorrect linking, we corrected it manually when possible, or discarded it. There can be sometimes a discrepancy between the information retrieved with Wikipedia and the role of the entity in the New York Time context, for entities whose role has changed greatly over the years. For instance, Donald Trump would have been considered in 2006 as a businessman, whereas in 2020, he's first of all president of the United-States. Yet, we did not try to account for this effect, as we figured it could be confusing for the annotators if the information displayed is in contradiction with the annotator's own knowledge. Besides, this effect is not major as it still makes sense to mention an entity in

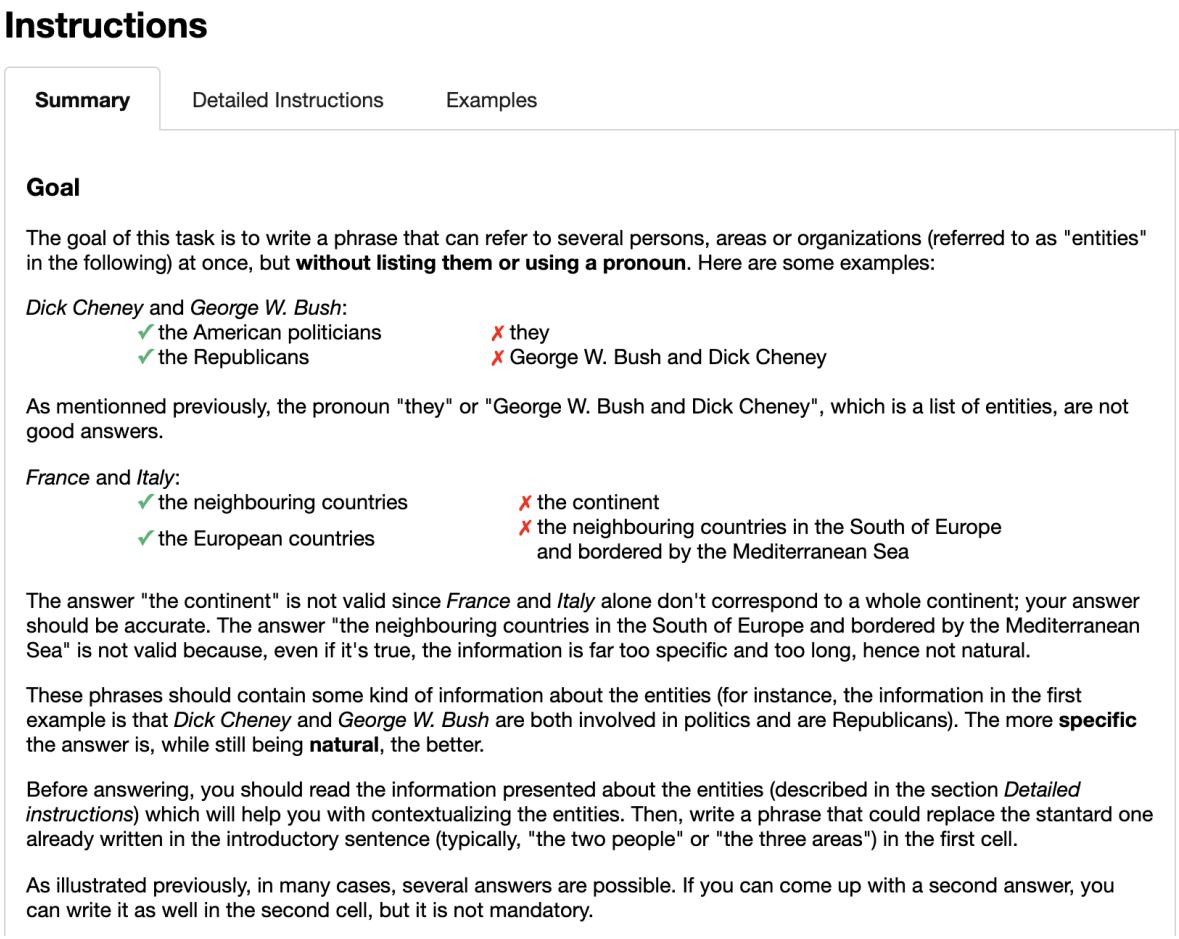
[1]Using `Wikipedia` python's library: https://pypi.org/project/wikipedia/

the context of its former occupation, and, after collecting the annotations, we did not notice any particular issue with it.

After extraction, we sampled 2,100 instances uniformly at random for annotation. A tuple contains between 2 and 6 entities, for an average of 2.4.

3.3 Data Annotation

We used Amazon Mechanical Turk to collect entity aggregations. Annotators were asked to generate aggregations given information about an aggregatable instance. For each instance, we showed the same information as described above, including the mentions of the entities in context, and a link to the Wikipedia pages of the entities. We present in details the layout of the annotation process in Figure 1, and its instructions in Figures 2, 3 and 4.



Instructions

Summary | Detailed Instructions | Examples

Goal

The goal of this task is to write a phrase that can refer to several persons, areas or organizations (referred to as "entities" in the following) at once, but **without listing them or using a pronoun**. Here are some examples:

Dick Cheney and *George W. Bush*:

✓ the American politicians
✓ the Republicans
✗ they
✗ George W. Bush and Dick Cheney

As mentionned previously, the pronoun "they" or "George W. Bush and Dick Cheney", which is a list of entities, are not good answers.

France and *Italy*:

✓ the neighbouring countries
✓ the European countries
✗ the continent
✗ the neighbouring countries in the South of Europe and bordered by the Mediterranean Sea

The answer "the continent" is not valid since *France* and *Italy* alone don't correspond to a whole continent; your answer should be accurate. The answer "the neighbouring countries in the South of Europe and bordered by the Mediterranean Sea" is not valid because, even if it's true, the information is far too specific and too long, hence not natural.

These phrases should contain some kind of information about the entities (for instance, the information in the first example is that *Dick Cheney* and *George W. Bush* are both involved in politics and are Republicans). The more **specific** the answer is, while still being **natural**, the better.

Before answering, you should read the information presented about the entities (described in the section *Detailed instructions*) which will help you with contextualizing the entities. Then, write a phrase that could replace the stantard one already written in the introductory sentence (typically, "the two people" or "the three areas") in the first cell.

As illustrated previously, in many cases, several answers are possible. If you can come up with a second answer, you can write it as well in the second cell, but it is not mandatory.

Figure 2: First page of the instructions provided to the annotators.

Instructions

Summary | **Detailed Instructions** | Examples

Full example

Title of the article: Iraq Findings Leaked by Aide Were Disputed
[...] The official refused to be named because he was not authorized to discuss the issue. Why those three men were acting so quietly remains a mystery, and Mr. Bush and **Mr. Cheney** have never discussed it in public. [...]

Dick Cheney	George W. Bush
Richard Bruce Cheney is an American politician and businessman who served as the 46th vice president of the United States from 2001 to 2009. He has been cited as the most powerful vice president in American history.	George Walker Bush is an American politician and businessman who served as the 43rd president of the United States from 2001 to 2009. He had previously served as the 46th governor of Texas from 1995 to 2000.

In this article, Dick Cheney and George W. Bush are discussed. The two people...
[Replace "The two people" with another phrase referring to the entities]

- ✓ the politicians
- ✓ the Republicans
- ✓ the American statesmen
- ✗ they
- ✗ George W. Bush and Dick Cheney

The entities

In this task, you are presented with several entities of the same type (either people, areas, or organizations). In the example presented above, the entities are people: *Dick Cheney* and *George W. Bush*. Your task is to replace the standard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally which is **natural** and brings information that is **specific** to the context of the entities' mentions.

Be careful, your answer **should not be a list of the entities, nor a pronoun**.

Information available

New York Times excerpt

Title of the article: Iraq Findings Leaked by Aide Were Disputed
[...] The official refused to be named because he was not authorized to discuss the issue. Why those three men were acting so quietly remains a mystery, and Mr. Bush and **Mr. Cheney** have never discussed it in public. [...]

As you can see above, you could think about several different phrases for the same entities, therefore having a few sentences to help you contextualize your answer can help. These sentences come from a *New York Times* article, and should be used to determine what information you wish to underline: ideally, your answer should be **specific** to the excerpt (this means that your answer could be used in the same context as the article).

In order to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which entity it is referring to). However, it is an automatic process, so in some cases there might be some bugs. In that case you should try to give an answer without taking the highlighting into account.

Wikipedia information

Dick Cheney	George W. Bush
Richard Bruce Cheney is an American politician and businessman who served as the 46th vice president of the United States from 2001 to 2009. He has been cited as the most powerful vice president in American history.	George Walker Bush is an American politician and businessman who served as the 43rd president of the United States from 2001 to 2009. He had previously served as the 46th governor of Texas from 1995 to 2000.

In order to write a relevant phrase, you also need to have some knowledge about the entities. In the example above, you cannot write any of the answers if you don't know that *Dick Cheney* and *George W. Bush* are politicians, Republicans, etc. You can use your own knowledge, but in case you don't know enough about the entities, we provide some information extracted from *Wikipedia*. If you need more information, the name of the entity is a link to the *Wikipedia* page.

Note that the *Wikipedia* information is just an help. Sometimes no information is found about an entity or the information retrieved is about an irrelevant article (espcially when namesakes can be found). In that case you should still try to give an answer, using only what you know or what you can infer from the example.

Issue

In some cases, it can happen that you may not be able to find an answer. Such situation can happen for several reasons: not enough information about the entities, no common ground or opposition between them, or bug in the example that you cannot overcome... In that case, you should write "N/A" (for "no answer") or in short "NA" in the first cell, and check the corresponding box.

Figure 3: Second page of the instructions provided to the annotators.

Instructions

Summary | Detailed Instructions | **Examples**

Good examples

Title of the article: Western Powers Disagree on Elements of Iran Proposal

[...] The disagreements on these issues are clouding the possibility of a deal with Iran on its nuclear program, even as tensions have increased over Tehran 's refusal to change its behavior, the diplomats said. In addition, they said, **Europe**, the United States and **Russia** have not agreed on the need to impose sanctions on Iran if it continues to defy the West. [...]

Europe	Iran	Russia
Europe is a continent located entirely in the Northern Hemisphere and mostly in the Eastern Hemisphere. It is bordered by the Arctic Ocean to the north, the Atlantic Ocean to the west, Asia to the east, and the Mediterranean Sea to the south. It comprises the westernmost part of Eurasia.	Iran, also called Persia, and officially the Islamic Republic of Iran, is a country in Western Asia. With 82 million inhabitants, Iran is the world's 18th most populous country. Its territory spans 1,648,195 km2 , making it the second largest country in the Middle East and the 17th largest in the world.	Russia , or the Russian Federation , is a transcontinental country in Eastern Europe and North Asia. At 17,125,200 square kilometres , Russia is, by a considerable margin, the largest country in the world by area, covering more than one-eighth of the Earth's inhabited land area.

In this article, Europe, Iran and Russia are discussed. The three areas...

This is a difficult example since the three entities don't have an obvious common ground (two of them are countries, the other one is a continent). Here are several answers you can consider:

- ✓ **the political powers**: these areas are either important countries or continents, therefore you can consider them as "political powers";
- ✓ **the conflicting entities**: considering the context established by the article, we can learn that these three areas are in some conflict (*Europe* and *Russia* are defied by *Iran*), therefore you can come up with an answer around this piece of information;
- ✗ **the three countries**: it is important not to give an answer that doesn't work for every entity; since Europe is not a country, "the three countries" cannot be accepted here.

Among the answers you can think about, remember that you should chose the one that seems to you the most specific while still being natural.

Bad examples

Title of the article: In the Race For Governor, A Big Divide On School Aid

[...] Many policy makers say the state's economic prospects hinge on improving schools. ''For New York State to succeed and become vibrant again, education and economic development have to become top priorities,'' said Abraham M. Lackman, who has been both a top budget aide to the Republican-controlled State Senate and a New York City budget director.'' [...]

New York City	New York State
The City of New York, usually called either New York City or simply New York , is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over a land area of about 302.6 square miles , New York is also the most densely populated major city in the United States.	New York is a state located in the Northeastern United States. New York was one of the original thirteen colonies that formed the United States. With an estimated 19.54 million residents in 2018, it is the fourth most populous state. In order to distinguish the state from its city with the same name, it is sometimes referred to as New York State .

In this article, New York City and New York State are discussed. The two areas...

This example can be considered as problematic since the two entities are not really distinct: *New York City* is located inside *New York State*. Here are a few answers you can consider:

- ✓ **N/A**: you can consider that no natural answer can be found since the entities are not really distinct;
- ✓ **the New York jurisdicitons**: if you can come up with an answer that overcomes the difficulty of the example, you can give it as well;
- ✓ **the New York region**: another way to find an answer is to find something that gather the entities, here the region can account for the city and the state;
- ✗ **the state and the city of New York**: remember that you should not list the entities (you can consider it as a list as soon as there is "and" in the answer);
- ✗ **the largest metropolitan area in the world by urban landmass and one of the world's most populous megacities**: even if this fact is true, it is by far too specific and really unnatural.

Again, your choice of answer should be based on the following question: which one is the most natural and specific? You should especially chose whether you want to give an answer or not ("N/A"), depending on if you think you could use naturally your answer, or if it's not natural at all.

Figure 4: Third page of the instructions provided to the annotators.

The entity tuple, document context, Wikipedia background information are presented to annotators, alongside a prompt (see Figure 1). For the PERSON entities in our example, this prompt is *"In this article, François Bayrou, Nicolas Sarkozy and Ségolène Royal are discussed. The three people..."* Annotators were asked to replace the phrase "The three people" with a relevant one referring to the entities. The prompt serves to prime the annotator to produce a fluent and comprehensive aggregation covering all the entities. For other named entity types, the prompt is changed accordingly. While simple, we found this prompt to be rather effective in the collection process.

We also presented detailed examples (see Figure 2) explaining the desired aggregations. Annotators were asked not to use generic aggregations involving only the entities' type (e.g., "the three people") and to avoid using "and", as it would often imply an enumeration.

For each of the 2,100 aggregatable instances, three different annotators were asked to provide an annotation. In each annotation, an annotator could provide between zero (meaning the instance is not aggregatable) and two aggregations. The unprocessed aggregations produced for the examples of Table 2 by the annotators are available in Table 3.

We discarded instances that at least two of the three annotators considered as 'not aggregatable'. In addition, we discarded those annotations that did not conform to our instructions, and annotations from workers who performed less than five annotations.

Finally, we post-processed the aggregations, removing determiners, numerical expressions and standardized the casing (e.g., "The two cities" became "cities").

Table 4 presents statistics on the size of the data collected and the final dataset.

Aggregatable instance	Aggregations
Input entities Francois Bayrou, Nicolas Sarkozy and Segolene Royal **Context** Street Violence by Paris Youths Intrudes Again Into French Politics: **The Socialist candidate, Ségolène Royal**, who is running second in the opinion polls, said the incident showed that **Mr. Sarkozy** had failed as interior minister." In five years with a right-wing government that has made crime its main campaign issue, you can see that it is a failure all the way," she said on Canal+ television. **François Bayrou, a centrist presidential candidate**, also took aim at Mr. **Sarkozy**, saying," It is very important to end this climate of perpetual confrontation between police and some citizens." © *2008 The New York Times Company, used with permission*	**Annotation 1** • french politicians **Annotation 2** • the French politicians • the French presidential candidates **Annotation 3** • the politicians
Input entities Chicago and London **Context** Virtually Cool: The author of the hour was Chris Anderson, who after the drinks entertained the crowd with a simulcast PowerPoint lecture on the topic of his new best seller," The Long Tail," which describes how the chokehold of mass culture is being loosened by the new Internet-enabled economics of niche culture and niche commerce. The party was sponsored in part by a small SoHo-based new-media company called Flavorpill, which produces free e-mail magazines and weekly event guides for New York, Los Angeles, San Francisco, **Chicago** and **London**. © *2008 The New York Times Company, used with permission*	**Annotation 1** • major metropolitan cities **Annotation 2** • Cities **Annotation 3** • the cities • the major cities
Input entities Microsoft Corp. and Sony Corp **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. © *2008 The New York Times Company, used with permission*	**Annotation 1** • The technology companies **Annotation 2** • multinational corporations **Annotation 3** • the multinational corporations

Table 3: Examples of the entity aggregations collected through crowd-sourcing. For each aggregatable instance, we gathered three annotations from different workers, who could give between zero and two aggregations each. For displaying purposes, the aggregatable instances have been simplified.

Data collected	
aggregatable instances	2100
annotators	63
annotations	6299
Preprocessed dataset	
aggregatable instances	1718
annotators	42
annotations	4675
PERSON entities tuples	941 (801)
LOCATION entities tuples	629 (412)
ORGANIZATION entities tuples	148 (123)
PERSON aggregations	2900 (951)
LOCATION aggregations	2041 (505)
ORGANIZATION aggregations	456 (239)

Table 4: Statistics on the sizes of the annotated data and of the final dataset. For entity tuples and aggregations, we indicate the total count of occurrences, and in parentheses the count of unique occurrences.

3.4 Data Splits

We split the dataset into training, validation, and test sets using a 2:1:1 ratio, resulting in 858/430/430 aggregatable instances in each set, respectively (corresponding to 20592/10320/10320 ranking candidates, respectively).

The entities in our dataset are quite diverse. In the validation and test sets, 29% and 30% of the aggregatable instances respectively have a set of input entities which do not overlap with entities in the training set at all. On average, each aggregatable instance has 2.7 different aggregations.

4 The TESA task

To leverage the data collected in TESA dataset, we built a modeling task upon it. We first describe our design choices for TESA as a task, and describe the metrics we used to monitor the performances.

4.1 Task Definition

We frame TESA as a ranking task where, given an aggregatable instance as input, models must rank a list of candidates according to their plausibility as an aggregation of the input entities in context. We choose a discriminative approach to avoid relying on word-overlap metrics (e.g. ROUGE-based metrics), and we opt for a ranking task set-up to avoid classification between heavily imbalanced classes, as the number of gold standards remains limited, facing the number of candidate aggregations. Besides, in this set-up, generative models can also be evaluated by computing their perplexity of a given a candidate aggregation.

In our experiments, the list of candidate aggregations is chosen to contain 24 candidates in total, including the gold-standard, correct aggregations generated by the human annotators, as well as a list of negative candidates which serve as distractors. The candidates' number is chosen to yield approximately 10 times more negative candidates than gold standards. Negative candidates are sampled uniformly at random from other aggregatable instances sharing the same named entity type, in order to make the task harder. An example of TESA's ranking tasks is available in Table 5.

BART-based models' input	Candidates to rank
François Bayrou is a French centrist politician [...], who was a candidate in the 2002, 2007 and 2012 French presidential elections. Nicolas Paul Stéphane Sarkozy [...] is a retired French politician who served as President of France [...] from 16 May 2007 until 15 May 2012. Ségolène Royal [...] is a French politician and former Socialist Party candidate for President of France. Street Violence by Paris Youths Intrudes Again Into French Politics: The Socialist candidate , Ségolène Royal , who is running second in the opinion polls, said the incident showed that Mr. Sarkozy had failed as interior minister. [...] François Bayrou , a centrist presidential candidate , also took aim at Mr. Sarkozy , saying," It is very important to end this climate of perpetual confrontation between police and some citizens." Francois Bayrou, Nicolas Sarkozy, Segolene Royal	afghans, police officers, **french presidential candidates**, intelligence analysts, tv talent, american lobbyists, former presidents, defectors, former boxers, **politicians**, real estate company owners, participants in anna nicole smith case, american men, **french politicians**, new york mafiosos, people involved in the scandal, iraqi citizens, billionaire businessmen, male speed skaters, investors, men involved in professional sports, screen artists, poets, alleged criminals
Chicago , locally also), officially the City of Chicago, is the most populous city in the U.S. state of Illinois and the third most populous city in the United States. With an estimated population of 2,705,994 , it is also the most populous city in the Midwestern United States. [...] London is the capital and largest city of England and the United Kingdom. Standing on the River Thames in the south-east of England, at the head of its 50-mile estuary leading to the North Sea, London has been a major settlement for two millennia. [...] Virtually Cool: The author of the hour was Chris Anderson, who after the drinks entertained the crowd with a simulcast PowerPoint lecture on the topic of his new best seller," The Long Tail," which describes how the chokehold of mass culture is being loosened by the new Internet-enabled economics of niche culture and niche commerce. The party was sponsored in part by a small SoHo-based new-media company called Flavorpill, which produces free e-mail magazines and weekly event guides for New York, Los Angeles, San Francisco, Chicago and London . Chicago, London	western asia cities, **major cities**, western-asia countries, eastern european locales, large political entities, neighboring middle eastern countries, rival nations, east coast states, major american cities, middle eastern counties, **major metropolitan cities**, eastern locations, african locations, central asian countries, sovereign states of the usa, security council members, new england areas, middle eastern regions, saudi arabian neighbors, places near the mediterranean sea, **cities**, iraqi areas, surrounding countries, political climates
Microsoft Corporation is an American multinational technology company with headquarters in Redmond, Washington. It develops, manufactures, licenses, supports, and sells computer software, consumer electronics, personal computers, and related services. Its best known software products are the Microsoft Windows line of operating systems, the Microsoft Office suite, and the Internet Explorer and Edge web browsers. [...] Sony Corporation is a Japanese multinational conglomerate corporation headquartered in Kōnan, Minato, Tokyo. Its diversified business includes consumer and professional electronics, gaming, entertainment and financial services. 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. Microsoft Corp., Sony Corp.	multinational consumer electronics corporations, militant groups, american entertainment companies, transportation organizations, entertainment groups, **technology companies**, palestinian political organizations, palestinian political parties, rivals, medical organizations, hockey teams, entities of the palestinian legislative council, multinational aerospace corporation, **multinational corporations**, communications groups, transportation corporations, business partners, military organizations, california organizations, retailers, new york city organizations, american pharmaceutical company, political organizations, european telecommunications firms

Table 5: Examples of TESA's ranking tasks. BART-based models' inputs are presented in the left-hand-side column. Background information is in blue, context is in violet, and entities' names are in orange. Models have to rank the 24 candidates (separated by commas) of the right-hand-side column. The gold standard aggregations are in bold. For displaying purposes, these examples have been shortened.

4.2 Evaluation Measures

We evaluate the models' performances using three widely used ranking performance measures. Let *rank*(i) be the rank of candidate i, G be the set of gold-standard candidates in a ranking and $R(n)$ be the set of candidates which are ranked between 1 (best rank) and n (n included). Then, for an aggregatable instance, we use the following metrics.

Average precision:

$$AP = \frac{1}{|G|} \sum_{i \in G} \frac{|G \cap R(\textit{rank}(i))|}{|R(\textit{rank}(i))|} \tag{1}$$

Recall at 10:

$$R@10 = \frac{1}{|G|} |G \cap R(10)| \tag{2}$$

Reciprocal rank:

$$RR = \frac{1}{\min_{i \in G} \textit{rank}(i)} \tag{3}$$

We report the mean of these values across all instances in the test set (MAP, $\overline{R@10}$, MRR). We chose these measures because they provide different perspectives on the evaluation results. For instance, the recall at 10 captures the models' ability to rank correct aggregations as promising or neutral at worst, whereas the reciprocal rank focuses solely on the best ranked correct aggregation. In the following, when comparing the models and their performances, we use the average precision.

5 Models

In order to evaluate the difficulty of our task and to give an example of how state-of-the-art models can be evaluated on our task, we tested several simple baselines as well as models adapted from current work on abstractive summarization on TESA. We tested BART (Lewis et al., 2019) as a representative model of recent high-performance abstractive summarization systems based on an encoder-decoder architecture with a Transformer backbone. We compared three versions of BART, which differ based on whether and how they are fine-tuned on TESA. For all three versions above, we built upon code that is available through fairseq (Ott et al., 2019). We use the version of BART pre-trained on the CNN/DailyMail dataset.

5.1 Simple Baselines

All the baselines and models are given as input an aggregatable instance and a list of candidates to rank with the same entity type as the aggregatable instance. The first two baselines are agnostic to the aggregatable instance:

Random. This baseline produces a random ordering of the candidate entities.

Frequency. This baseline ranks the candidates according to their frequency as a correct aggregation in the training set.

5.2 Logistic Regression

5.2.1 Model

We defined a number of statistical and linguistically informed features, which we extracted from each candidate aggregation and its aggregatable instance's context and background information. We then trained from scratch a binary logistic regression using this representation, to discriminate between the gold-standard aggregations and the negative candidates. At evaluation time, we used the score attributed by the model for each candidate to establish the ranking. More precisely, we ranked the candidates according to the likelihood, given by the model, of belonging to the class

"gold standard" (the higher the score, the better the rank, or, in other words, the closer to 1).

5.2.2 Linguistically informed features

The 15 features we used for this model are:

- count of the "frequency" baseline,
- number of common tokens (with repetition) between the candidate and the union of the entities' background information,
- size of the word overlap between a candidate and the union of the entities' background information,
- size of the word overlap between a candidate and the intersection of the entities' background information,
- number of entities whose background information words are overlapping the candidate's words,
- cosine similarity between the average token embeddings of the candidate and the union of the entities' background information,
- cosine similarity between the average word embeddings of the candidate and the intersection of the entities' background information,
- number of common tokens (with repetition) between the candidate and the context,
- size of the word overlap between a candidate and the context,
- cosine similarity between the average token embeddings of the candidate and the context,
- number of common tokens (with repetition) between the candidate and the union of the entities' background information, the context, and the names of the entities,
- size of the word overlap between a candidate and the union of the entities' background information, the context, and the names of the entities,
- size of the word overlap between a candidate and the union of the context and the intersection of the entities' background information,
- cosine similarity between the average token embeddings of the candidate and the union of the entities' background information, the context, and the names of the entities,
- cosine similarity between the average word embeddings of the candidate and the union of

the context and the intersection of the entities' background information.

During the feature extraction, we removed any capitalization and any punctuation. We removed the stop-words from the candidates' tokens. We removed the stop-words and we lemmatized the tokens of the context and of the background information.

5.2.3 Hyperparameters

We used a simple logistic regression for binary classification. The model has 32 parameters, and we use Adam optimizer, a learning rate of $3e-3$ and the cross entropy loss. We ran the experiment for 50 epochs, which took typically 15 minutes on a CPU, and we kept the model's parameters of the epoch maximizing the average precision of the validation set.

5.3 Pre-trained BART

5.3.1 Model

We applied an existing pre-trained version of BART in a generative set-up without fine-tuning. We formatted each aggregatable instance into a single sequence of tokens by concatenating the fields of the aggregatable instances in the following order: background information, context (title of the article and excerpt), and entity names. An example of such input can be seen in Table 5.

We fed this as input to BART's encoder, and we evaluated the probability of each candidate aggregation to be generated autoregressively by the decoder. We used these probabilities to rank the candidates.

5.3.2 Hyperparameters

To evaluate pre-trained BART, we used the following parameters to evaluate candidates' likelihood:

- `beam=10`,
- `lenpen=1.0`,
- `max_len_b=100`,
- `min_len=1`,

- `no_repeat_ngram_size=2`.

This model had 401 million parameters, none of them was trained in this approach.

5.4 Generative BART

5.4.1 Model

The generative BART version is similar to the above, pre-trained BART, but this time, we fine-tuned BART on TESA, considering each correct aggregation as a separate target, and training the model to generate each target given the corresponding aggregatable instance as input. For the aggregatable instances, we used the same input format as above. We did not add any form of separation tokens, as our initial experiments showed that they (slightly) hurt the performance.

5.4.2 Hyperparameters

To finetune generative BART, our choice of hyperparameter search and final hyperparameters was inspired by BART's finetuning on summarization datasets described here. We kept the model's parameters of the experiment and the epoch maximizing the average precision of the validation set. We performed a grid search on the following hyperparameters:

- `lr` in `{3e-6, 5e-6, 1e-5, 2e-5, 3e-5}`,
- `max-tokens` in `{1024, 2048}`.

We used the following final hyperparameters:

- `lr=5e-06`,
- `max-tokens=1024`,
- `max-epochs=6`,
- `update-freq=1`,
- `total-num-updates=4974`,
- `warmup-updates=149`.

`total-num-updates` was determined empirically as $\frac{\texttt{max-epochs} \cdot \texttt{updates-per-epoch}}{\texttt{update-freq}}$ and `warmup-updates` was chosen as 3% of `total-num-updates`. During the hyperparameter search we used

`total-num-updates=4974, 375` and `warmup-updates=149, 67` for `max-tokens=1024, 2048` respectively.

To evaluate candidates' likelihood and to generate aggregations, we modified slightly the code of Ott et al. (2019) to compute all hypotheses of the beam search (not only the most probable one) and we used the same parameters as in Section 5.3.2. We ran our experiments on a single V100 GPU with 32GB of memory with the `fp16` option, and an experiment took typically 1 hour. This model had 401 million parameters, all of them being trained.

For the ablation study, we used the final hyperparameters, except for `total-num-updates` and `warmup-updates` which were determined empirically as above. We added `max-sentences=16` for the "entities" ablation experiment.

5.5 Discriminative BART

5.5.1 Model

Finally, we fine-tuned BART discriminatively as a classifier. During fine-tuning, we consider each candidate and its aggregatable instance as a separate sample, and the model was trained on these samples to discriminate the correct aggregations from the negative candidates. At test time, we rank the candidates by their probability of being the correct aggregation according to the classifier. Again, we did not add any separation tokens, as it did not improve the performance.

The main advantage of this approach over the previous one is that it leverages the set-up of TESA as a ranking task, and the model is exposed to both correct and incorrect aggregations during training (which, on the other hand, makes it more computationally expensive). By contrast, generative BART only sees correct ones. We thus expect the discriminative model to produce higher performance. However, this comes at a cost, as this approach cannot generate freely an aggregation, but only retrieve one from a set of candidates.

5.5.2 Hyperparameters

For this approach, our choice of hyperparameter search and final hyperparameters was largely inspired by BART's finetuning on GLUE tasks (Wang et al., 2018) described here. We kept the

model's parameters of the experiment and the epoch maximizing the average precision of the validation set. We performed a grid search on the following hyperparameters:

- `lr` in `{5e-6, 1e-5, 2e-5, 3e-5}`,
- `max-sentences` in `{4, 8, 16}`.

We used the following final hyperparameters:

- `lr=2e-5`,
- `max-sentences=8`,
- `num-classes=2`,
- `max-epochs=6`,
- `total-num-updates=18180`,
- `warmup-updates=1090`.

`total-num-updates` was determined empirically as $\frac{\texttt{max-epochs} \cdot \texttt{updates-per-epoch}}{\texttt{update-freq}}$ and `warmup-updates` was chosen as 6% of `total-num-updates`. During the hyperparameter search we used `total-num-updates=30888, 18180, 16254` and `warmup-updates=1853, 1090, 975` for `max-sentences=4, 8, 16` respectively. We ran each experiment on a single V100 GPU with 32GB of memory with the `memory-efficient-fp16` option, and an experiment took typically 5 hours. This model had 401 million parameters, all of them being trained.

For the ablation study described in Chapter 6.1, we used the following hyperparameters, as they yielded very similar performances:

- `lr=1e-5`,
- `max-tokens=1024`,
- `max-sentences=8`,
- `update-freq=4`,
- `max-epochs=5`.

`total-num-updates` and `warmup-updates` were determined empirically as above.

6 Results

The results of the models described in Chapter 5 on TESA's test set are presented in Table 6. We see that most models outperform the frequency baseline, except for pre-trained BART. Fine-tuning BART on TESA increased its performance significantly, especially if done discriminatively. Discriminative BART achieves the best results. Its high performance can be mitigated by our choice of ranking only 24 candidates, which makes less likely confusing negative candidates.

For reproducibility purposes, we include in Table 7 the validation scores corresponding to the main results of Table 6.

6.1 Ablation Study

To understand the importance of the different components of the input for this task, we performed an ablation study, where we removed selected parts of the input: without the background information (context, entities), without context (info., entities) and with only the names of the entities (entities). We fine-tuned generative and discriminative BART on these modified datasets.

We report the mean average precision results, which are representative of the other measures, in Table 8. Models perform best when all information is available, which validates our choice of input format. The background information seems to be more important than the context, as removing the context leads to the smallest drop in average precision. Interestingly, models perform quite well when given only the entities' names, though the performance gap is still quite significant.

Method	MAP	$\overline{R@10}$	MRR
Random baseline	0.222	0.442	0.289
Frequency baseline	0.570	0.655	0.761
Logistic regression	0.700	0.863	0.840
Pre-trained BART	0.389	0.682	0.505
Generative BART	0.701	0.903	0.840
Discriminative BART	**0.895**	**0.991**	**0.954**

Table 6: Results of the different models on the TESA test set.

Method	MAP	$\overline{R@10}$	MRR
Random baseline	0.226	0.415	0.304
Frequency baseline	0.557	0.637	0.773
Logistic regression	0.675	0.843	0.834
Pre-trained BART	0.385	0.666	0.488
Generative BART	0.684	0.882	0.835
Discriminative BART	**0.892**	**0.980**	**0.964**

Table 7: Results of the different models on the TESA validation set.

Method	context, entities	info., entities	entities
Generative BART (0.701)	-0.079	-0.049	-0.145
Discriminative BART (0.895)	-0.035	-0.024	-0.100

Table 8: Results of the ablation study. We report the mean average precision differences between the ablated system and the full model's performance (in parentheses) on TESA. Negative numbers mean the performance of the model without ablation is higher.

6.2 Qualitative analysis

We compare the two best-performing models: generative and discriminative BART. In Table 9, we present a few examples of their results on a ranking task from TESA's test set. In general, the discriminative approach performs well, is robust and the negative candidates ranked at high positions are quite coherent (e.g., from Table 9, "former presidents" and "police officers"). On the other hand, generative BART performs quite well on the ranking task, but is far less robust and its negative candidates ranked at high positions are more intriguing (e.g., from Table 9, "new york mafiosos" and "american men"), which seems to indicate a poorer understanding of the aggregatable instance.

Besides, we show some aggregations generated by the generative approach in Table 10. Qualitatively speaking, the generated samples are quite interesting as many of them are accurate and have a diverse vocabulary (e.g., from Table 10, "politicians", "figures", "candidates", "leader"). However, some samples are factually inconsistent (e.g., from Table 10, "american politicians") which seems to indicate that the model does not have a deep understanding of relevant semantic concepts (e.g., nationalities cannot be substituted for each other).

In Tables 11 and 12, we display some examples specifically chosen as the models failed on them. The high performances of the models demonstrate that these kind of examples are a minority, however, we want to exhibit them to show what kind of examples can make our models fail. Note that in these cases, the examples are difficult to solve even for the human judgement, as they are both somehow flawed: relying on noisy data from TESA dataset or containing a confusing false negatives.

Discriminative BART	Generative BART
Entities Francois Bayrou, Nicolas Sarkozy and Segolene Royal	
1. **politicians** 2. **french politicians** 3. **french presidential candidates** 4. former presidents 5. police officers 6. alleged criminals	1. **politicians** 2. **french politicians** 3. people involved in the scandal 4. **french presidential candidates** 5. new york mafiosos 6. american men
Entities Chicago and London	
1. **cities [0.993]** 2. **major cities [0.980]** 3. **major metropolitan cities [0.970]** 4. major american cities [0.149] 5. new england areas [0.031] 6. political climates [0.008]	1. **cities [0.067]** 2. major american cities [0.049] 3. neighboring middle eastern countries [0.038] 4. eastern european locales [0.036] 5. **major cities [0.034]** 6. surrounding countries [0.022] 10. **major metropolitan cities [0.016]**
Entities Microsoft Corp. and Sony Corp.	
1. **technology companies [0.988]** 2. **multinational corporations [0.951]** 3. multinational consumer electronics corporations [0.899] 4. business partners [0.029] 5. rivals [0.022] 6. communications groups [0.001]	1. **multinational corporations [0.063]** 2. **technology companies [0.056]** 3. multinational consumer electronics corporations [0.039] 4. american entertainment companies [0.036] 5. entertainment groups [0.028] 6. retailers [0.019]

Table 9: Results of generative and discriminative BART on the running examples. We show the input entities, and the candidates ranked from 1 to 6, as well as any other gold standard candidate, if any. Gold standards are in bold; the candidates' likelihoods predicted by the models are in brackets.

Entities François Bayrou, Nicolas Sarkozy and Ségolène Royal	**Entities** Chicago and London	**Entities** Microsoft Corp. and Sony Corp.
1. **politicians [0.084]** 2. american politicians [0.060] 3. **french politicians [0.057]** 4. political figures [0.041] 5. **French politicians [0.037]** 6. political leaders [0.029] 7. politician [0.025] 8. political candidates [0.024] 9. politicans [0.023] 10. politicians [0.008]	1. american cities [0.087] 2. **cities [0.067]** 3. political powers [0.054] 4. american regions [0.045] 5. american areas [0.044] 6. **major cities [0.034]** 7. politicians [0.030] 8. us cities [0.027] 9. world cities [0.026] 10. people [0.009]	1. multinational companies [0.067] 2. corporations [0.065] 3. **multinational corporations [0.063]** 4. american companies [0.057] 5. **technology companies [0.056]** 6. tech companies [0.049] 7. companies [0.040] 8. businesses [0.034] 9. countries [0.032] 10. technology firms [0.028]

Table 10: Aggregations generated by generative BART on the running examples. The model's encoder is fed an aggregatable instance, and the decoder generates autoregressivly the aggregations without constraint. We show the input entities, and the 10 aggregations retrieved by the beam search, ranked according to their likelihoods. If a generated aggregation matches a gold standard (except for capital letters), it is in bold; the generated examples probabilities are in brackets.

BART-based models' input	Discriminative BART	Generative BART
Cobra Verde is a 1987 German drama film directed by Werner Herzog and starring Klaus Kinski, in their fifth and final collaboration. [...] Klaus Kinski was a German actor.He appeared in more than 130 films, and was a leading role actor in the films of Werner Herzog, including [...] Cobra Verde . [...] Where Heart of Darkness Begets Head of Nuttiness: Along with" Aguirre" and" Fitzcarraldo,"" Cobra Verde" completes a trilogy of mayhem and megalomania in hot climates. Mr. Kinski is the title character , a Brazilian rancher , originally known as Francisco Manoel da Silva, who turns to banditry after being driven from his land by drought and famine. Cobra Verde, Klaus Kinski	1. **german [0.742]** 2. **aspects of the german film world [0.323]** 3. companions [0.156] 4. parties involved [0.006] 5. show business professionals [0.001] 6. contributors [0.001]	1. contributors [0.047] 2. people with an interest in politics [0.032] 3. **aspects of the german film world [0.029]** 4. singer-songwriters [0.025] 5. political figures [0.019] 6. mafiosi [0.019] 23. **german [0.002]**
After 40 Years, 2 Hotel Plans Vie for Port Washington's Heart: The Bradley is awaiting a zoning variance and site plan approval from the Town of North Hempstead and could start construction next summer, Mr. D'Alonzo said. Mr. D'Alonzo and his partner, Sam Suzuki of the real estate company Vintage Group , said they had met several times with local officials and residents and, in response to those comments, agreed to reduce the number of rooms to 46 and lower the building 's height to 40 feet. Joe D'Alonzo, Sam Suzuki	1. developers [0.989] 2. **partners [0.982]** 3. **real estate company owners [0.940]** 4. **businessmen [0.921]** 5. pair [0.161] 6. washington-area residents [0.102]	1. **real estate company owners [0.050]** 2. **businessmen [0.037]** 3. developers [0.026] 4. coworkers [0.025] 5. **partners [0.020]** 6. american investors [0.020]

Table 11: Examples of TESA's ranking tasks which were poorly solved by generative and discriminative BART. We show the candidates ranked from 1 to 6, as well as any other gold standard candidate, if any. Gold standards are in bold; the candidates' likelihoods predicted by the models are in brackets. For displaying purposes, these examples have been shortened. Both examples can be considered as noisy and difficult to solve, as they could fool human judgement: in the first example the set of entities is made of a person and a movie; in the second example, the candidate "developers" is relevant to the aggregatable instance and can be considered as a false negative.

Entities Cobra Verde, Klaus Kinski	**Entities** Joe D'Alonzo and Sam Suzuki
1. entertainers [0.091]	1. hotel owners [0.091]
2. filmmakers [0.079]	2. hotel developers [0.083]
3. american actors [0.067]	3. Hotel owners [0.071]
4. film industry professionals [0.063]	4. hotel plans [0.070]
5. american filmmakers [0.051]	5. Hotel developers [0.069]
6. politicians [0.049]	6. hotel partners [0.067]
7. German film actors [0.048]	7. Hotel partners [0.059]
8. actors [0.046]	8. hotels [0.053]
9. directors [0.042]	9. **businessmen [0.037]**
10. film makers [0.042]	10. business partners [0.036]

Table 12: Examples of the aggregations generated by generative BART on the examples of Table 11. We show the input entities, and the 10 aggregations retrieved by the beam search, ranked according to their likelihoods. If a generated aggregation matches a gold standard (except for capital letters), it is in bold; the generated examples probabilities are in brackets.

7 Conclusion and future work

In this thesis, we propose TESA, a novel task and an accompanying dataset of crowd-sourced entity aggregations in context. TESA directly measures the ability of summarizers to abstract at a semantic level, on a particular task: the entity aggregation. It contains aggregations that are document-specific and have a high quality, thanks to the annotation process we have performed. TESA is the main contribution of our work, and we hope it can be useful for many researchers, in order to test their models on our dataset and our task.

In this work, we also compare several baseline models and models adapted from existing abstractive summarizers on TESA, as a proof of concept. We first establish that TESA is an interesting task, inasmuch as it is neither trivial, nor impossible to solve for models. Besides, we find that a discriminative fine-tuning approach achieves the best performance, though this model inherently cannot *generate* aggregations, besides being heavily more expensive, computationally speaking. Its generative counterpart remains also interesting, despite having slightly poorer performances, inasmuch as it is much easier to use in real-world applications. For both approaches, qualitatively speaking, the aggregations retrieved are satisfying.

In future work, we would like to expand the domains covered by our dataset, which is biased towards topics found in the source corpus, such as politics for the New York Time corpus. This bias could be problematic when applying a model trained only on TESA to a real-world dataset, but it could be reduced easily by exploring data from other domains. Another important direction is to investigate how to integrate the ability to aggregate entities derived from training on TESA into an abstractive summarizer. The skills of summarization and entity aggregation seems difficult to merge together, inasmuch as our task is difficult to integrate to a summarization dataset in an end-to-end fashion. However, having one model for summarization and one model for entity aggregation seems to be a simpler approach, yet, this would require models to tackle another challenging issue which we have not addressed and which cannot be answered with our dataset: which set of entities *should* a model aggregate in the first place?

References

Nitin Agarwal, Ravi Shankar Reddy, Kiran Gvr, and Carolyn Penstein Rosé. 2011. Towards multi-document summarization of scientific articles:making interesting comparisons with SciSumm. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pages 8–15, Portland, Oregon. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg. Springer-Verlag.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

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).

Meng Cao and Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. 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 3163–3172, Hong Kong, China. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.

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.

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

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.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *Computing Research Repository*, arXiv:1910.12840.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising equence-to-sequence pre-training for natural language generation, translation, and comprehension. *Computing Research Repository*, arXiv:1910.13461.

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.

Elena Lloret, María Teresa Romá-Ferri, and Manuel Palomar. 2013. Editorial: Compendium: A text summarization system for generating abstracts of research papers. *Data Knowl. Eng.*, 88:164–175.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. 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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times annotated corpus. In *Philadelphia: Linguistic Data Consortium. LDC2008T19. DVD*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 697–706, New York, NY, USA. Association for Computing Machinery.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 19–27, USA. IEEE Computer Society.