TESA: A Task in Entity Semantic Aggregation for Abstractive Summarization

Clément Jumel McGill University, Mila clement.jumel@mail.mcgill.ca Annie Louis Google Research, London annielouis@google.com

Jackie C. K. Cheung

McGill University, Mila jcheung@cs.mcgill.ca

Abstract

Human-written texts contain frequent generalizations and semantic aggregation of content. 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" and "two European cities". Yet generation, especially, abstractive summarization systems have so far focused heavily on paraphrasing and simplifying the source content, to the exclusion of such semantic abstraction capabilities. In this paper, we present a new dataset and task aimed at the semantic aggregation of entities. TESA contains a dataset of 5.3K crowd-sourced entity aggregations of PERSON, ORGANIZATION, and LO-CATION named entities.¹ 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 finetune on TESA an encoder-decoder language model 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.

1 Introduction

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. Concretely, recent summarization datasets including XSum (Narayan

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	

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.

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.

 $^{^1}TESA's$ code and access to its dataset can be found at: <code>https://github.com/clementjumel/tesa</code>

In this paper, we start to tackle this issue by focusing on the specific task of semantic aggregation of entities; i.e., how to refer to a tuple of named entities using a noun phrase instead of enumerating them (See Table 1 for an example). We define a task to evaluate summarization models on semantic entity aggregation, which we call TESA (A Task in Entity Semantic Aggregation). 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 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, despite some factual inconsistencies.

2 Related work

Abstractive summarizers have gained prominence with the popularization of RNNs (Sutskever et al., 2014; Nallapati et al., 2016), and more recently Transformers (Vaswani et al., 2017) like BERT (Devlin et al., 2019). Several abstractive models have achieved state-of-the-art performances on benchmark summarization datasets in terms of ROUGE, including ProphetNet (Yan et al., 2020), PEGA-SUS (Zhang et al., 2019) and BART (Lewis et al., 2019). 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).

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.

Our proposed entity aggregation task is related to referring expression generation (REG). 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. Our work is also related to coreference resolution, especially those that examine multi-antecedent resolution (Burga et al., 2016; Vala et al., 2016), an inverse problem to ours. To the best of our knowledge, no previous work has directly addressed entity aggregation.

3 The TESA dataset

We used the New York Times (NYT) Annotated Corpus (Sandhaus, 2008) to extract tuples of named entities and their document context. The NYT corpus contains high-quality metadata listing the salient named entities mentioned in each article. We form our tuples from entities tagged in the metadata for the same article.

We refer to a tuple of entities and its associated information as an **aggregatable instance**. We

first describe the components of an aggregatable instance in more detail. Then, we describe our data extraction and crowd-sourcing experiments.

3.1 An aggregatable instance

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, is in Figure 1. For more examples, see Table 8 in the appendix.

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 sets of entities.

We used the 2006 and 2007 portions of the New York Times 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

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 2: 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.

its Wikipedia page², and, in case of ambiguous or incorrect linking, we corrected it manually when possible, or discarded 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. Some of the instructions given to the annotators and examples of the annotation layout are in Figures 1, 2. The complete instructions and examples are available in Figures 3–6 in the appendix.

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

²Using Wikipedia python's library: https://pypi. org/project/wikipedia/

View instructions

[] The Socialist candidate, Ségolène Royal, who is running right-wing government that has made crime its main campaigr	treet Violence by Paris Youths Intrudes Again Into French second in the opinion polls, said the incident showed that Mr. Sarkozy had f issue, you can see that it is a failure all the way," she said on Canal+ televisi zy, saying," It is very important to end this climate of perpetual confrontation	ailed as interior minister.'' In five years with a on. François Bayrou,		
Francois Bayrou Nicolas Sarkozy Segolene Royal				
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 Paul Stéphane Sarközy de Nagy-Bocsa ; born 28 January 1955) is a retired French politician who served as President of France and exorption of the Democratic Movement and former socialist Party candidate for President of Presiden				
In this article, Francois Bayrou, Nicolas Sarkozy and Segolene Royal are discussed. The three people Replace "The three people" with your most relevant phrase referring to the entities				
If you have a second answer, you can write it here				
Issues: • remember that you can't answer "They" or an	y pronoun	Subm		
 if you cannot find any relevant phrase mentionning all the entities, write "NA" in the first cell above and check the following box: I can't find a relevant phrase 				

Figure 1: Layout of the annotation task. The mentions of the entities in the New York Times article are colored and the name of the corresponding entity is visible when an annotator clicks on a mention. The title of the Wikipedia information is an hyperlink to the corresponding web page.

structio		
Summary	Detailed Instructions	Examples
Goal		
		at can refer to several persons, areas or organizations (referred to as "entities' og them or using a pronoun. Here are some examples:
Dick Cheney a	nd George W. Bush:	
	the American politicians	× they
\checkmark	the Republicans	X George W. Bush and Dick Cheney
As mentionne good answers	1 37 1	ney" or "George W. Bush and Dick Cheney", which is a list of entities, are not
France and Ita	ıly:	
\checkmark	the neighbouring countries	× the continent
~	the European countries	X the neighbouring countries in the South of Europe and bordered by the Mediterranean Sea
hould be acc	urate. The answer "the neigh	ce <i>France</i> and <i>Italy</i> alone don't correspond to a whole continent; your answern bouring countries in the South of Europe and bordered by the Mediterranear the information is far too specific and too long, hence not natural.
example is that		of information about the entities (for instance, the information in the first <i>N. Bush</i> are both involved in politics and are Republicans). The more specific better.
nstructions) w	which will help you with conte	prmation presented about the entities (described in the section <i>Detailed</i> extualizing the entities. Then, write a phrase that could replace the stantard or e (typically, "the two people" or "the three areas") in the first cell.
		everal answers are possible. If you can come up with a second answer, you

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

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 aggregations produced for the example of Figure 1 by the three annotators are below:

Annotator 1

• french politicians

Annotator 2

- the French politicians
- the French presidential candidates

Annotator 3

• the politicians

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

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

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, and we opt for a ranking task set-up to avoid classification between heavily imbalanced classes, as the number of gold standards remains limited. In this set-up, generative models can also be evaluated.

In our experiments, the list of candidate aggregations contains 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.

An example of TESA's tasks is available in Table 3; for more examples, see Table 9 in the appendix.

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 retrieved up to and including position n. Then, for an aggregatable instance:

Average precision.

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

Recall at 10.

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

Reciprocal rank.

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

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. Recall

BART-based models' input

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

Candidates to rank

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

Table 3: Ranking task from the running example. 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, this example has been shortened.

at 10 captures the models' ability to rank correct aggregations as promising or neutral at worst. Reciprocal rank focuses solely on the best ranked correct aggregation.

5 Models

We tested several simple baselines as well as models adapted from current work on abstractive summarization on TESA.

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

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. These 15 features include:

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

We detail these features in Appendix C. We trained a binary logistic regression using this representation, to discriminate between the gold-standard aggregations and the negative candidates. We used the model's predictive probability for the gold-standard class to produce a ranking over the candidate list.

5.3 BART-based models

We tested BART (Lewis et al., 2019) as a representative model of recent high-performance abstractive summarization systems based on an encoderdecoder architecture with a Transformer backbone. We compared three versions of BART, which differ based on whether and how they are fine-tuned on TESA.

Pre-trained BART. We applied an existing pretrained 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 3.

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.

Generative BART. This version is similar to the above, but 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. 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.

Discriminative BART. Finally, we fine-tuned BART discriminatively as a classifier. During finetuning, 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.

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. The choice of hyperparameters is described in Appendix D.

6 Results

The results of the models on TESA's test set are presented in Table 4. We see that most models out-

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 4: Results of the different models on the TESAtest set.

Method	context, entities	info., entities	entities
	cilities	cintities	<u> </u>
Generative BART (0.701)	-0.079	-0.049	-0.145
Discriminative BART (0.895)	-0.035	-0.024	-0.100

Table 5: 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 full model is higher.

perform the frequency baseline, except 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 unlikely confusing negative candidates.

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. The hyperparameters used are described in Appendix D.

We report the mean average precision results, which are representative of the other measures, in Table 5. 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.

Discriminative BART	Generative BART
Entities Francois Ba and Segolene Royal	yrou, Nicolas Sarkozy
 politicians french politicians french presidential candidates former presidents police officers alleged criminals 	 politicians french politicians people involved in the scandal french presiden- tial candidates new york mafiosos american men

Table 6: Results of generative and discriminative BART on the running example. 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.

A	4 1
Aggregations	generated
11551 CSaulons	Scheratea

Entities François Bayrou, Nicolas Sarkozy and Ségolène Royal

- 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]

Table 7: Aggregations generated by generative BART on the running example. 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.

6.2 Qualitative analysis

We compare the two best-performing models: generative and discriminative BART. In Table 6, we present an example 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., "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., "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 7. Qualitatively speaking, the generated samples are quite interesting as many of them are accurate and have a diverse vocabulary (e.g., "politicians", "figures", "candidates", "leader"). However, some samples are factually inconsistent (e.g., "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).

For other examples, including some specifically chosen as the models failed on them, see Tables 10–13 in the appendix.

7 Conclusion and future work

We have proposed 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. We have compared several baseline models and models adapted from existing abstractive summarizers on TESA, and find that a discriminative fine-tuning achieves the best performance, though this model inherently cannot *generate* aggregations.

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. Another important direction is to investigate how to integrate the ability to aggregate entities derived from training on TESA into an abstractive summarizer. This would require models to tackle another challenging issue which we have not addressed: which set of entities *should* a model aggregate in the first place?

Acknowledgements

We would like to thank Ido Dagan for useful initial discussions about this work, Jingyun Liu, whose automatic preprocessing of the New York Time Corpus was a precious help, and José Ángel González Barba, who helped us with analysis and proofreading. We also wish to mention our two pilot annotators. Finally, we would like to thank our anonymous reviewers for their kind and helpful comments.

This work was supported by the Natural Sciences and Engineering Research Council of Canada, and the last author is supported by a Canada CIFAR AI Chair.

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

Appendix A Detailed examples

In Tables 8, 9, 10 and 11, we present several examples following the examples presented in Figure 1 and Tables 3, 6 and 7 respectively.

In Tables 12 and 13, we present two examples of aggregatable instances where BART-based models performed poorly.

Appendix B Human-sourcing set-up

To ensure maximal reproducibility, we provide here some details regarding the collection of the human annotations. For the task set-up, we used the Amazon Mechanical Turk website. We present in details the layout of the annotation process in Figure 3, and its instructions in Figures 4, 5 and 6.

Appendix C Linguistically informed features

For the representation of an aggregatable instance to train the logistic regression, we used the following features:

- 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.

Appendix D Hyperparameters

In the following, we describe our choice of hyperparameters for each model, as well as any eventual hyperparameter search.

D.1 Logistic regression

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.

D.2 Pre-trained BART

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.

D.3 Generative BART

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 max-epochs-updates-per-epoch update-free and warmup-updates chosen as was 3% of total-num-updates. Durthe hyperparameter search used ing we total-num-updates=4974, 375 warmup-updates=149, 67 for and 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 Appendix D.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.

D.4 Discriminative BART

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 empirmax-epochs.updates-per-epoch ically as update-free and warmup-updates was chosen as 6% of total-num-updates. During hyperparameter search used the we total-num-updates=30888, 18180, and 16254 warmup-updates=1853, 1090, 975 max-sentences=4, 8, for 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, 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.

Appendix E Validation results

For reproducibility purposes, we include in Table 14 the validation scores corresponding to the main results, in Table 4.

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**. © 2008 The New York Times Company, used with permission

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. [©] 2008 The New York Times Company, used with permission

Table 8: Examples of aggregatable instances and their crowd-sourced aggregations. 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 (underlined title and excerpt with mentions of the entities in bold). For each aggregatable instance, we gathered three annotations from different workers, who could give between zero and two aggregations each. For displaying purposes, these examples have been shortened.

Annotation 1

• major metropolitan cities

Annotation 2

• Cities

Annotation 3

- the cities
- the major cities

Annotation 1

• The technology companies

Annotation 2

• multinational corporations

Annotation 3

• the multinational corporations

BART-based models' input

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

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 Konan, 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.

Candidates to rank

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

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 9: Ranking tasks from the running examples. 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.

Discriminative BART	Generative BART
Entities Chicago and London	
 cities [0.993] major cities [0.980] major metropolitan cities [0.970] major american cities [0.149] new england areas [0.031] political climates [0.008] 	 cities [0.067] major american cities [0.049] neighboring middle eastern countries [0.038] eastern european locales [0.036] major cities [0.034] surrounding countries [0.022] major metropolitan cities [0.016]
Entities Microsoft Corp. and Sony Corp. 1. technology companies [0.988]	1. multinational corporations [0.063]
2. multinational corporations [0.951]	2. technology companies [0.056]
 multinational consumer electronics corpora- tions [0.899] 	 3. multinational consumer electronics corpora tions [0.039]
4. business partners [0.029]	4. american entertainment companies [0.036]
5. rivals [0.022]	5. entertainment groups [0.028]
6. communications groups [0.001]	6. retailers [0.019]

Table 10: 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 Chicago and London	Entities Microsoft Corp. and Sony Corp.
1. american cities [0.087]	1. multinational companies [0.067]
2. cities [0.067]	2. corporations [0.065]
3. political powers [0.054]	3. multinational corporations [0.063]
4. american regions [0.045]	4. american companies [0.057]
5. american areas [0.044]	5. technology companies [0.056]
6. major cities [0.034]	6. tech companies [0.049]
7. politicians [0.030]	7. companies [0.040]
8. us cities [0.027]	8. businesses [0.034]
9. world cities [0.026]	9. countries [0.032]
10. people [0.009]	10. technology firms [0.028]

Table 11: 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 star- ring 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, in- cluding [] Cobra Verde . [] Where Heart of Darkness Begets Head of Nut- tiness: Along with" Aguirre" and" Fitz- carraldo,"" Cobra Verde" completes a trilogy of mayhem and megalomania in hot climates. Mr. Kinski is the title char- acter , 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	 german [0.742] aspects of the german film world [0.323] companions [0.156] parties involved [0.006] show business professionals [0.001] contributors [0.001] 	 contributors [0.047] people with an interest in politics [0.032] aspects of the german film world [0.029] singer-songwriters [0.025] political figures [0.019] mafiosi [0.019] 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 construc- tion next summer, Mr. D'Alonzo said. Mr. D'Alonzo and his partner, Sam Suzuki of the real estate company Vin- tage 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	 developers [0.989] partners [0.982] real estate company owners [0.940] businessmen [0.921] pair [0.161] washington-area residents [0.102] 	 real estate company owners [0.050] businessmen [0.037] developers [0.026] coworkers [0.025] partners [0.020] american investors [0.020]

Table 12: 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 13: Examples of the aggregations generated by generative BART on the examples of Table 12. 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.

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 14: Validation results of the different models on TESA, for reproducibility purposes.

View instructions

ght-wing government that has made crime its main campaigr	second in the opinion polls, said the incident showed that Mr. Sarkozy had fi issue, you can see that it is a failure all the way," she said on Canal- televisic zy, saying," It is very important to end this climate of perpetual confrontation	on. François Bayrou,
Francois Bayrou	Nicolas Sarkozy	Segolene Royal
rançois Bayrou is a French centrist politician and the resident of the Democratic Movement , who was a andidate in the 2002, 2007 and 2012 French presidential lections.	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.	Ségolène Royal ; born 22 September 1953), is a French politician and former Socialist Party candidate for President of France.
In this article, Francois Bayrou, Nicolas Sarkozy and Segolen Replace "The three people" with your most relevant phrase ref		
If you have a second answer, you can write it here		
Issues: • remember that you can't answer "They" or an	y pronoun	Subn
 if you cannot find any relevant phrase mention 	ning all the entities, write "NA" in the first cell above and check the following l	box:

Figure 3: Layout of the annotation task. The mentions of the entities in the New York Times article are colored and the name of the corresponding entity is visible when an annotator clicks on a mention. The title of the Wikipedia information is an hyperlink to the corresponding web page.

structio	ons	
Summary	Detailed Instructions	Examples
Goal		
		at can refer to several persons, areas or organizations (referred to as "entities" ig them or using a pronoun . Here are some examples:
✓	nd <i>George W. Bush:</i> the American politicians	<mark>⊁</mark> they
V	the Republicans	🗡 George W. Bush and Dick Cheney
As mentionned good answers	1 27 1	ney" or "George W. Bush and Dick Cheney", which is a list of entities, are not
France and Ita	5	
V	the neighbouring countries	X the continent
√	the European countries	* the neighbouring countries in the South of Europe and bordered by the Mediterranean Sea
should be acc	urate. The answer "the neigh	ce <i>France</i> and <i>Italy</i> alone don't correspond to a whole continent; your answer nbouring countries in the South of Europe and bordered by the Mediterranean the information is far too specific and too long, hence not natural.
example is tha		of information about the entities (for instance, the information in the first <i>N. Bush</i> are both involved in politics and are Republicans). The more specific better.
instructions) w	hich will help you with conte	prmation presented about the entities (described in the section <i>Detailed</i> extualizing the entities. Then, write a phrase that could replace the stantard on e (typically, "the two people" or "the three areas") in the first cell.
	previously, in many cases, se well in the second cell, but i	everal answers are possible. If you can come up with a second answer, you

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

Instructions

natural and brings information that is specific to the context of the entities' mentions. e careful, your answer should not be a list of the entities, nor a pronoun. Information available ew York Times excerpt Title of the article: Iraq Findings Leaked by Alde 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. [] s you can see above, you could think about several different phrases for the same entities, therefore having a few mentences to help you contextualize your answer can help. These sentences come from a New York Times article, and hould be used to determine what information you wish to underline: ideally, your answer should be specific to the corpt (this means that your answer could be used in the same context as the article). order to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which enti-				
Image: Control of the state of the stat	ull examp	le		
Image in the politician and the several entries in public. [] Dick Cheney George W. Bush Richard Bruce Cheney is an American politician and businessman who served as the 48th vice president of the United States from 2001 2008. He had previously served as the 48th governor of Texas president in American history. George Walker Bush is an American politician and businessman who served as the 48th governor of Texas president in American history. 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 American statesmen the American statesmen the American statesmen the another phrase that can refer to all the entities at once, and ideally which natural and brings information that is specific to the context of the entities mentions. e careful, your answer should not be a list of the entities, nor a pronoun. formation available ew York Times excerpt The official refused to be named because he was not authorized to discuss the issue. Why those three men were acting so quiety remains a mystery, and Mr. Bush and Mr. Cheney have never discussed it in public. [] you can see above, you could think about several different phrases for the same entities, therefore having a few meneres to help you contextualize your answer can help. These sentences come from a New York Times excile, and you answer should be specific to the cortext at the artacide).		Title of the	e article: Iraq Findings	s Leaked by Aide Were Disputed
Image in the politician and the several entries in public. [] Dick Cheney George W. Bush Richard Bruce Cheney is an American politician and businessman who served as the 48th vice president of the United States from 2001 2008. He had previously served as the 48th governor of Texas president in American history. George Walker Bush is an American politician and businessman who served as the 48th governor of Texas president in American history. 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 American statesmen the American statesmen the American statesmen the another phrase that can refer to all the entities at once, and ideally which natural and brings information that is specific to the context of the entities mentions. e careful, your answer should not be a list of the entities, nor a pronoun. formation available ew York Times excerpt The official refused to be named because he was not authorized to discuss the issue. Why those three men were acting so quiety remains a mystery, and Mr. Bush and Mr. Cheney have never discussed it in public. [] you can see above, you could think about several different phrases for the same entities, therefore having a few meneres to help you contextualize your answer can help. These sentences come from a New York Times excile, and you answer should be specific to the cortext at the artacide).	() The (0)			
Bichard Bruce Cheney is an American politician and businessman who served as the 46th vice 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 (Peopleace "The two people" with another phrase referring to the entities] * the politicians * they * the Republicans * they * the American statesmen * George W. Bush and Dick Cheney * the Republicans * they * the American statesmen * George W. Bush and Dick Cheney * the American statesmen * George W. Bush and Dick Cheney * the American statesmen * they * the American statesmen * George W. Bush. Your task is to replace the andrard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally which ratural and brings information that is specific to the context of the entities "mentions. e careful, your answer should not be a list of the entities, nor a pronoun. Afformation available * 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. Cheney have never discussed ti in public [] * you can see above, you could think about several different phrases for the same entities, therefore having a few menteres to help you contextualize youranswer can help. These sentences come from a New York Times artic				
Bichard Bruce Cheney is an American politician and businessman who served as the 46th vice 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 (Peopleace "The two people" with another phrase referring to the entities] * the politicians * they * the Republicans * they * the American statesmen * George W. Bush and Dick Cheney * the Republicans * they * the American statesmen * George W. Bush and Dick Cheney * the American statesmen * George W. Bush and Dick Cheney * the American statesmen * they * the American statesmen * George W. Bush. Your task is to replace the andrard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally which ratural and brings information that is specific to the context of the entities "mentions. e careful, your answer should not be a list of the entities, nor a pronoun. Afformation available * 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. Cheney have never discussed ti in public [] * you can see above, you could think about several different phrases for the same entities, therefore having a few menteres to help you contextualize youranswer can help. These sentences come from a New York Times artic				
who served as the 46th vice president of the United States from 2001 who served as the 47th United States from 2001 2001 to 2009. He has been cited as the most powerful vice b 2009. He had previously served as the 46th governor of Texas for 2009. He had previously served as the 46th governor of Texas for 2009. He had previously served as the 46th governor of Texas for 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] * they * the Republicans * they * the American statesmen * George W. Bush and Dick Cheney this task, you are presented with several entities of the same type (either people, areas, or organizations). In the cample presented above, the entities are people: Dick Cheney and George W. Bush. Your task is to replace the andard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally which natural and brings information that is specific to the context of the entities" mentions. e careful, your answer should not be a list of the entities, nor a pronoun. formation available ew York Times excerpt YHe of ficial 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. [] syou can see above, you could think about several different phrases for the same entities, therefore		Dick Cheney		George W. Bush
who served as the 46th vice president of the United States from 2001 who served as the 46th greek on the determinant 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 [Peplace "The two people" with another phrase referring to the entities] 				
president in American history. from 1995 to 2000. In this article, Dick Cheney and George W. Bush are discussed. The two people [<i>Replace "The two people" with another phrase referring to the entities</i>]				
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]			powerful vice	
[Replace "The two people" with another phrase referring to the entities] • ' the politicians X they • ' the Republicans X George W. Bush and Dick Cheney • ' the American statesmen X George W. Bush and Dick Cheney this task, you are presented with several entities of the same type (either people, areas, or organizations). In the cample presented above, the entities are people: Dick Cheney and George W. Bush. Your task is to replace the andard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally which natural and brings information that is specific to the context of the entities' mentions. e careful, your answer should not be a list of the entities, nor a pronoun. Information available ew York Times excerpt The official refused to be named because he was not authorized to discuss the issue. Why those three men were acting so quietly means a mystery, and Mr. Bush and Mr. Cheney have never discussed it in public. [] s you can see above, you could think about several different phrases for the same entities, therefore having a few entences to help you contextualize your answer can help. These sentences come from a New York Times article, and yould be used to determine what information you wish to underline: ideally, your answer should be specific to the coerpt (this means that your answer could be used in the same context as the article).				
				
• ' the Republicans • ' the American statesmen * Ceorge W. Bush and Dick Cheney • the American statesmen * Ceorge W. Bush and Dick Cheney * the American statesmen * Ceorge W. Bush and Dick Cheney * the entities this task, you are presented with several entities of the same type (either people, areas, or organizations). In the rample presented above, the entities are people: <i>Dick Cheney</i> and <i>George W. Bush</i> . Your task is to replace the andard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally which natural and brings information that is specific to the context of the entities' mentions. • careful, your answer should not be a list of the entities, nor a pronoun. • formation available • w York Times excerpt	[Replace "	The two people" with anothe	er phrase referring	to the entities]
If the American statesmen the entities this task, you are presented with several entities of the same type (either people, areas, or organizations). In the cample presented above, the entities are people: <i>Dick Cheney</i> and <i>George W. Bush.</i> Your task is to replace the andard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally whice natural and brings information that is specific to the context of the entities' mentions. e careful, your answer should not be a list of the entities, nor a pronoun. Information available ew York Times excerpt Take of the article. Iraq Findings Leaked by Alde 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. [] s you can see above, you could think about several different phrases for the same entities, therefore having a few mentences to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and hould be used to determine what information you wish to underline: ideally, your answer should be specific to the cortext as the article). order to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which entities)	√	the politicians		× they
this task, you are presented with several entities of the same type (either people, areas, or organizations). In the tample presented above, the entities are people: <i>Dick Cheney</i> and <i>George W. Bush</i> . Your task is to replace the andard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally whice natural and brings information that is specific to the context of the entities' mentions. e careful, your answer should not be a list of the entities, nor a pronoun . formation available ew 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. [] a you can see above, you could think about several different phrases for the same entities, therefore having a few metherces to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and tould be used to determine what information you wish to underline: ideally, your answer should be specific to the corept (this means that your answer could be used in the same context as the article). order to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which ent				X George W. Bush and Dick Cheney
this task, you are presented with several entities of the same type (either people, areas, or organizations). In the ample presented above, the entities are people: <i>Dick Cheney</i> and <i>George W. Bush</i> . Your task is to replace the andard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally whic natural and brings information that is specific to the context of the entities' mentions. a careful, your answer should not be a list of the entities, nor a pronoun . formation available ew York Times excerpt Title of the enticies 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. 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 meners to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and ould be used to determine what information you wish to underline: ideally, your answer should be specific to the carticle to the larger to the same context as the article).	√	the American statesmen		
ample presented above, the entities are people: <i>Dick Cheney</i> and <i>George W. Bush.</i> Your task is to replace the andard phrase (here: "The two people") with another phrase that can refer to all the entities at once, and ideally whic natural and brings information that is specific to the context of the entities' mentions. e careful, your answer should not be a list of the entities, nor a pronoun . formation available aw York Times excerpt Title of the articles traq 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. [] a you can see above, you could think about several different phrases for the same entities, therefore having a few metheoses to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and ould be used to determine what information you wish to underline: ideally, your answer should be specific to the carticle, and ould be used to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which entities).	ne entities	3		
Information available we York Times excerpt Title of the articles 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. [] syou can see above, you could think about several different phrases for the same entities, therefore having a few intences to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and hould be used to determine what information you wish to underline: ideally, your answer should be specific to the corpt (this means that your answer could be used in the same context as the article).	ample prese andard phra	ented above, the entities are se (here: "The two people")	people: Dick Cher with another phras	ney and George W. Bush. Your task is to replace the e that can refer to all the entities at once, and ideally whic
Information available we York Times excerpt Title of the articles 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. [] s you can see above, you could think about several different phrases for the same entities, therefore having a few nuences to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and hould be used to determine what information you wish to underline: ideally, your answer should be specific to the corpt (this means that your answer could be used in the same context as the article).				
Sew York Times excerpt Title of the article: Iraq Findings Leaked by Alde 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. [] a you can see above, you could think about several different phrases for the same entities, therefore having a few intences to help you contextualize your answer can help. These sentences come from a New York Times article, and ould be used to determine what information you wish to underline: ideally, your answer should be specific to the cerpt (this means that your answer could be used in the same context as the article). order to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which entities have been colored in the excerpt (you can click on a colored word to see to which entities have been colored in the excerpt (you can click on a colored word to see to which entities have been colored in the excerpt (you can click on a colored word to see to which entities have been colored in the excerpt (you can click on a colored word to see to which entities the article).	e careiui, yo	ur answer should not be a l	ist of the entities,	nor a pronoun.
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. [] syou can see above, you could think about several different phrases for the same entities, therefore having a few interces to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and ould be used to determine what information you wish to underline: ideally, your answer should be specific to the cerpt (this means that your answer could be used in the same context as the article).	formation	n available		
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. [] syou can see above, you could think about several different phrases for the same entities, therefore having a few interces to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and ould be used to determine what information you wish to underline: ideally, your answer should be specific to the cerpt (this means that your answer could be used in the same context as the article).	w Vork Tin	nes excernt		
[] 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. [] Is you can see above, you could think about several different phrases for the same entities, therefore having a few intences to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and would be used to determine what information you wish to underline: ideally, your answer should be specific to the corpt (this means that your answer could be used in the same context as the article).				
[] 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. [] s you can see above, you could think about several different phrases for the same entities, therefore having a few intences to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and nould be used to determine what information you wish to underline: ideally, your answer should be specific to the corpt (this means that your answer could be used in the same context as the article).		Title of the	article: Irao Findings	Leaked by Aide Were Disputed
remains a mystery, and Mr. Bush and Mr. Cheney have never discussed it in public. [] s you can see above, you could think about several different phrases for the same entities, therefore having a few intences to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and hould be used to determine what information you wish to underline: ideally, your answer should be specific to the ccerpt (this means that your answer could be used in the same context as the article).		The of the	and the second se	
s you can see above, you could think about several different phrases for the same entities, therefore having a few entences to help you contextualize your answer can help. These sentences come from a <i>New York Times</i> article, and nould be used to determine what information you wish to underline: ideally, your answer should be specific to the coerpt (this means that your answer could be used in the same context as the article).				
entences to help you contextualize your answer can help. These sentences come from a New York Times article, and nould be used to determine what information you wish to underline: ideally, your answer should be specific to the coerpt (this means that your answer could be used in the same context as the article). order to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which ent	remains a n	nystery, and Mr. Bush and Mr. Che	eney have never discus	sed it in public. []
entences to help you contextualize your answer can help. These sentences come from a New York Times article, and nould be used to determine what information you wish to underline: ideally, your answer should be specific to the coerpt (this means that your answer could be used in the same context as the article). order to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which ent		a alaassa suassaassid thials ala		t - human fau tha anna antitica thaufau having a fau
ccerpt (this means that your answer could be used in the same context as the article). order to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which enti				
order to help you, the entities have been colored in the excerpt (you can click on a colored word to see to which ent				
is referring to). However, it is an automatic process, so in some cases there might be some bugs. In that case you				
nould try to give an answer without taking the highlighting into account.	order to hel			
and try to give an anomer manout taking the highlighting into account.	order to hel s referring t	o). However, it is an automat	ic process, so in s	ome cases there might be some bugs. In that case you

Wikipedia information

Dick Cheney	George W. Bush
Richard Bruce Cheney is an American politician and businessman	George Walker Bush is an American politician and businessman
who served as the 46th vice president of the United States from	who served as the 43rd president of the United States from 2001
2001 to 2009. He has been cited as the most powerful vice	to 2009. He had previously served as the 46th governor of Texas
president in American history.	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 5: Second page of the instructions provided to the annotators.

Instructions

ood examples			Bad examples
			Title of the entrole: In the Race For Governor, A Big Divide On School Aid
[] The disagreements deal with Iran on its nu over Tehran 's refusal t addition, they said, Eu agreed on the need to	Vestern Powers Disagree Proposal on these issues are cloud clear program, even as te o change its behavior, the rope, the United States ar impose sanctions on iran	ling the possibility of a nsions have increased diplomats said. In nd Russia have not	[] 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 West. []			The City of New York, usually called either New York City or New York City or
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 Arctic Ocean to 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.	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. In this article, New York City and New York State discussed. The two areas This example can be considered as problematic sin the two entities are not really distinct: <i>New York City</i> located inside <i>New York State</i> . Here are a few answ you can consider:
reas his is a difficult example ommon ground (two of t ere are several answers • √ the political pov	ran and Russia are disc since the three entities them are countries, the of you can consider: wers: these areas are eli fore you can consider t	don't have an obvious other one is a continen ther important countrie	 ✓ N/A: you can consider that no natural answ can be found since the entities are not really distinct; ✓ the New York jurisdicitons: if you can cor up with an answer that overcomes the difficul the example, you can give it as well; ✓ the New York region: another way to find
the article, we can (<i>Europe</i> and <i>Russia</i> with an answer aro • X the three countr doesn't work for ev three countries" ca mong the answers you	entities: considering the learn that these three ar a re defied by <i>Iran</i>), the und this piece of inform fies: it is important not to very entity; since Europe nnot be accepted here. can think about, rememin u the most specific while	eas are in some conflic refore you can come u ation; o give an answer that is not a country, "the per that you should che	 answer is to find something that gather the entities, here the region can account for the cand the state; X the state and the city of New York: rement that you should not list the entities (you can consider it as a list as soon as there is "and" the answer); X the largest metropolitan area in the workd's mappulous megacities: even if this fact is true 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 ar specific? You should especially chose whether you want to give an answer or not ("NA"), depending on you think you could use naturally your answer, or fit

Figure 6: Third page of the instructions provided to the annotators.