Summarizing Relationships for Interactive Concept Map Browsers

Abram Handler,[⋆] Prem Ganeshkumar,[†] Brendan O'Connor[⋆] and Mohamed AlTantawy[†]

Agolo[†]
New York, NY

College of Information and Computer Sciences[⋆]
University of Massachusetts, Amherst
ahandler@cs.umass.edu

Abstract

Concept maps are visual summaries, structured as directed graphs: important concepts from a dataset are displayed as vertexes, and edges between vertexes show natural language descriptions of the relationships between the concepts on the map. Thus far, preliminary attempts at automatically creating concept maps have focused on building static summaries. However, in interactive settings, users will need to dynamically investigate particular relationships between pairs of concepts. For instance, a historian using a concept map browser might decide to investigate the relationship between two politicians in a news archive. We present a model which responds to such queries by returning one or more short, importance-ranked, natural language descriptions of the relationship between two requested concepts, for display in a visual interface. Our model is trained on a new public dataset, collected for this task.

Code and data are available at:
https://github.com/slanglab/concept_maps_newsum19

1 Introduction

Concept maps are visual summaries, structured as directed graphs (Figure 1). Important concepts from a corpus are shown as vertexes. Natural language descriptions of the relationships between concepts are shown as textual labels, along the edges on the map. Initial attempts to generate English-language concept maps within natural language processing (Falke and Gurevych, 2017) have focused on creating static diagrams which summarize collections of documents.

However, in interactive settings, users will want to query relationships with a concept map interface, rather than simply read over fixed output from a summarization system. For instance, in the concept map browser shown in Figure 1, a user has queried for the relationship between Milosevic and Kosovo. An interactive system should include both concepts in a visual network, along with a labeled edge that summarizes their relationship (e.g. "Slobodan Milosevic revoked autonomy of Kosovo Province").

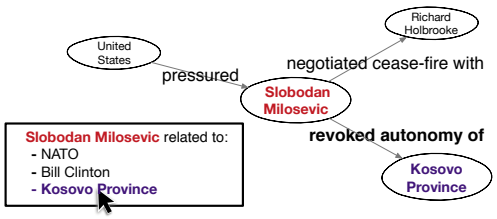

Figure 1: An example concept map browser. The system indicates that (t_1)="Slobodan Milosevic" is related to (t_2)="Kosovo Province." The user clicks to investigate the relationship, and the system must generate a summary explaining how Milosevic is related to Kosovo.

This study is concerned with how to add such labeled summary edges to a map. Given a pair of input query concepts, denoted (t_1) and (t_2), we attempt to select the best extractive, natural language *summary statement* which summarizes their relationship. Because there is no existing supervision to guide such a selection, we collect a new dataset of annotated summary statements, which we use to supervise a new model for this task.

Our study thus presents a full system for summarizing the relationship between an arbitrary pair of query concepts, extending prior work on relational summarization and concept maps (Falke and Gurevych, 2017; Handler and O'Connor, 2018).

2 Related work: relationship extraction

This study builds on prior efforts from Handler and O'Connor (2018), who propose extractively summarizing relationships via a two-stage process that first (1) identifies wellformed spans from a

Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 111–115

corpus that start with (t_1) and end with (t_2) and then (2) chooses the best summary statement from among these wellformed candidates. Handler and O'Connor (2018) show that extracting wellformed spans can find many more readable candidates than traditional relation extraction techniques. But they do not offer a method for the second step of picking a summary statement, which is the focus of this study.

We approach this new task of choosing the best summary statement from available candidates by collecting new supervision, tailored to the particular problem of summarizing relationships on concept maps. This form of supervision has a different focus from the existing Falke and Gurevych (2017) concept map dataset. Where Falke and Gurevych (2017) seek to create the best overall concept map for a given *topic*, this work seeks to find the best summary relationship for a given *relationship*. Therefore, unlike Falke and Gurevych (2017), our dataset includes labels for the most readable and informative statement describing the relationship between a $(t_1) - (t_2)$ query pair.

3 Overall technical approach

Like Handler and O'Connor (2018), we approach the problem of finding a short relationship *summary statement* with a two-stage approach.

Stage 1: We identify candidate summary statements using Handler and O'Connor (2018)'s method, which returns the probability that a span of tokens beginning with (t_1) and ending with (t_2) reads as a fluid and coherent sentence when extracted from naturally-occurring text.[1] (For brevity, we refer the reader to prior work for details, including discussion of why span extraction is preferred to relation extraction techniques). Table 1 provides examples of spans that do and do not make sense when extracted in this manner. We define all spans between (t_1) and (t_2) with a probability of well-formedness greater than .5 to be the **candidate set** for the pair (t_1) – (t_2). A sample candidate set is shown in Table 2.

Stage 2: In stage two, we choose the best summary statement from the candidate set. We collect new annotation to supervise this decision. Our annotation procedure assigns a score $\alpha(s) \in \{-3, -2, ...+3\}$ to each s in a candidate set,

[1]We also allow statements which begin with (t_2) and end with (t_1); the order of query concepts is important in interfaces which display concept maps, but beyond the scope of this work. We limit statements to a max. of 75 characters.

Milosevic withdrew from Kosovo ~~in 1999~~.
~~Clinton spoke with~~ Milosevic about Kosovo.

Table 1: Some spans (top) are plausible summary statements, because they make sense when removed from context sentences. Others spans (bottom) are not plausible summary statements because they don't make sense when extracted from sentences. We use an approach from Handler and O'Connor (2018) to identify such spans.

which is intended to reflect how well s summarizes a particular relationship. We use this supervision to train a model to predict $\alpha(s)$. We propose that the statement with the highest predicted $\alpha(s)$ score should be displayed on a concept map.

4 Candidate extraction

We approach the problem of summarizing relationships for concept maps by collecting a new dataset of annotated summary statements, drawn from news stories focusing on the Balkan Peninsula in the 1990s. Political scientists use rich news archives from this complex period to better understand conflict (Schrodt et al., 2001).

We create our dataset from *New York Times* articles (Sandhaus, 2008) published from 1990–1999, which mention at least one country from the Balkans. Following prior work on relational summarization, for each country, we use the package `phrasemachine` (Handler et al., 2016) to identify the 100 highest-frequency noun phrases within articles which mention that country.[2] The `phrasemachine` package uses a regular expression over part of speech tags to efficiently extract noun phrases, a useful syntactic category which includes both named entity spans (e.g. Boris Yeltsin) as well as other concepts (e.g. peace treaty). From all non-empty pairs of highest-frequency concepts, we sample a total of 689 pairs with more than two extracted candidates. In total there are 5,214 candidate statements across 689 sampled sets.[3] On average there are 7.56 state-

[2]https://github.com/slanglab/phrasemachine

[3]**Additional notes.** The countries are: Kosovo, Albania, Serbia, Croatia, Montenegro, Macedonia, Bulgaria, Romania, Moldova and Bosnia. (We exclude the former Yugoslavia; its landmass included other countries on our corpus). `phrasemachine` sometimes returns overlapping phrases, leading to duplicate sets. We merge duplicates with a heuristic which uses hand-written rules based on (i) token overlap between concepts and (ii) overlapping sentences be-

	A1	A2	A3
s_1 **General Grachev**'s favor is his loyalty to **Mr. Yeltsin**	-	W	-
s_2 **Mr. Yeltsin** openly accused **General Grachev**	-	-	-
s_3 **General Grachev**, Defense Minister by dint of his loyalty to **Mr. Yeltsin**	W	-	W
s_4 **General Grachev**'s plea today will do nothing to help **Mr. Yeltsin**	-	-	-
s_5 **Mr. Yeltsin** might also appear weak if he had to replace **General Grachev**	B	B	B

Table 2: A candidate set for (t_1) = "**Mr. Yeltsin**" and (t_2) = "**General Grachev**," along with decisions from three annotators (A1, A2 and A3) selecting the best (B) and worse (W) summary statement in the set. All annotators agree that s_5 is the best, so $\alpha(s_5)$ =3. (During annotation, the order of all sets was randomized).

ments per set ($\sigma = 10.6$).

5 Candidate annotation

5.1 Method

Some candidate sets in our dataset are easy for a person to judge and rank. For instance, it is possible to quickly read over the small set shown in Table 2 and identify statements which are clearly better and clearly worse synopses of the relationship between "General Grachev" and "Mr. Yeltsin".

However, other candidate sets in our dataset are too large and too complex to read and analyze quickly. (The largest candidate set in our dataset contains 143 statements in total). We accommodate both large and small sets with a "low-context" (Falke and Gurevych, 2017) annotation technique. We split candidate sets into one or more subsets, and ask annotators to rank the best and worst summary statements in each subset. Then we aggregate these local judgements about the best and worst candidates within each subset to create a global score. This global score, $\alpha(s)$, attempts to capture the overall quality of a given summary statement s.

This method of soliciting local judgements about subsets and then aggregating into an overall score is known as Best-Worst Scaling (Louviere, 1991). Best-Worst Scaling has been shown to make more efficient use of human judgements for a natural language task than traditional techniques (Kiritchenko and Mohammad, 2017).

5.2 Details of Best–Worst annotation

We present all candidate sets to three different non-native English speakers, hired via a professional annotation firm. All annotators completed graduate work in either linguistics or the humanities, and were based in the Middle East. For each annotator, we divide each candidate set into J random tuples (a tuple consists of up to eight candidate statements), and ask the annotator to choose the best and worst from each tuple. Annotators are instructed that the best statement should be the one that both sounds the most natural and that most helps them understand the history and politics of the Balkan region. They are instructed that the most unnatural sounding and least informative statement should be chosen as worst. In total, each candidate statement is shown to each annotator exactly once.[4] After annotators have judged each individual set, we aggregate with Orme (2009)'s counting formula: we set the score $\alpha(s) \in \{-3, -2, \ldots, +3\}$ of each summary statement s to be the number of times s was chosen as the best, minus the number of times it was chosen as the worst.

Following prior work (Kiritchenko and Mohammad, 2017), we evaluate inter-annotator agreement via split-half reliability. For each candidate set, we randomly split annotators into two groups, and compute the score for each s using each group of annotators. Then we compute the Spearman correlation (ρ) between the two sets of scores, yielding an average of $\rho = 0.495$ across 1000 random splits.

6 Modeling

The previous section describes a procedure for assigning a score, $\alpha(s)$ for each s in our dataset. We use these scores to train a model, $p(\alpha(s)|s)$. During modeling, we divide the dataset into training and test sets at the entity level, ensuring that there

tween sets. We exclude pairs which are very obviously unrelated to the Balkans (e.g. *Chinatown* and *Little Italy*). Our annotation budget determined the number of annotated sets.

[4] Unlike in traditional Best-Worst annotation, the number of candidates in each tuple may vary depending on the size of the candidate set. If a candidate set has a cardinality of less than eight, the size of the tuple is set to the size of the candidate set; otherwise the size of a tuple is capped at eight. We make this choice because many candidate sets have a small cardinality, and it does not make sense to break up small sets (e.g. 5 or 6 candidates) into very small tuples.

are no relationships between concepts in the training and test set. Ensuring that there are no relationships shared across sets is important because a model might use knowledge about relationships gleaned from training data (e.g. Milosevic led Serbia) to make inferences about relationships in the test data (e.g. Milosevic led the Serbian Socialist party). 627 candidates are used for training; the remaining 62 are for testing.[5]

We model $p(\alpha(s)|s)$ using ordinal regression, implemented with the MORD package (Pedregosa-Izquierdo, 2015). We use unigram features, morphological features, part-of-speech-tag features and binary features (e.g. s includes punctuation mark) to represent the candidate statement. Handler and O'Connor (2018)'s method (§4) returns a probability that a summary statement is grammatically wellformed. We include this probability as a feature in our model. We also include the token length of a summary statement as a feature. We tune MORD's regularization penalty parameter to maximize 5-fold, cross-validated Spearman's ρ using the training set.[6]

6.1 Evaluation and analysis

We use the test set to measure the extent to which our model's predictions correlate with gold scores, achieving a Spearman's $\rho = 0.443$ between our model's predictions and the gold scores. This is close to the $\rho = 0.495$ computed to measure inter-annotator agreement (§5.2).

We instructed annotators to select summary statements that were both informative and grammatically wellformed. We use the probability of grammatical well-formedness from the candidate detection method (§4) as a feature in our model. This measure appears to partially reflect annotator judgements: there is a Spearman's $\rho = 0.154$ between the two metrics across the dataset. Research into human perceptions of grammatical well-formedness (Sprouse and Schütze, 2014; Warstadt et al., 2018) could be applied to make better predictions in the future.

Model	Spearman's ρ
$\boldsymbol{p(\alpha(s)\|s)}$ (Ordinal regression)	**0.443**
Logistic regression	0.304
Inter-annotator agreement	0.495

Table 3: Spearman's ρ for our ordinal regression model $p(\alpha(s)|s)$, compared both to the inter-annotator agreement and a simpler logistic regression model.

Predicting annotator perceptions of informativeness is more challenging. For instance, annotators preferred "Mr. Milosevic has been formally charged with war crimes" $(\alpha(s) = 3)$ to "President Slobodan Milosevic may be indicted for war crimes" $(\alpha(s) = 1)$. The former expresses a completed action which arguably entails the latter, hypothetical action. How to best model (Bowman et al., 2015), formalize (MacCartney and Manning, 2009) and even study (Gururangan et al., 2018) such complex semantic relationships is an unsolved problem in NLP.

We use the number of tokens in a summary statement (subtracting out the length of query concepts) as a feature. We observe a Spearman's $\rho = .337$ between $\alpha(s)$ and the token length of s. We hypothesize that this feature might serve as a very coarse proxy for informativeness: although not instructed to do so, annotators might choose longer statements ahead of shorter statements because they express more about the Balkans.

7 Conclusion

We extend prior work focused on finding candidate summary statements (Handler and O'Connor, 2018) and constructing concept maps for an overall topic (Falke and Gurevych, 2017), by presenting a complete system for summarizing the relationship between an arbitrary pair of query concepts. Our method learns a model for selecting statements that best summarize relationships, which is supervised with a new, annotated resource for the task. We find that shallow cues like statement length and grammatical wellformedness are helpful for identifying good summary statements, but also that representing deeper semantic relationships (e.g. entailment) remains an ongoing challenge for automatically building concept maps.

Our study adopts the standard supervised paradigm underlying much current work on sum-

[5]To implement the train–test split, we form an initial provisional division of concepts into two sets. For all relationships between concepts that cross the two sets, we move the entity from the test set to the training set. All scored summary statements between concepts in the training set are used for training; the remainder are for test. We manually tune the size of the initial split so that 10% of concepts are in the final test set.

[6]We examine 10^i for $i = -3, -2..2, 3$ and use 10^1. Additionally, the MORD API implements several variants of ordinal regression. We use the LogisticSE variant because it achieves the highest cross-validated ρ on the training set.

marization (Hermann et al., 2015; Grusky et al., 2018). We gather human judgements of salience and well-formedness (in our case, judgements are expressed via Best-Worst Scaling), and then train a model to best replicate such judgements. Because such supervision is costly and difficult to collect, carries risks of annotation artifacts (Gururangan et al., 2018) and might transfer poorly to new domains, in the future, we plan to explore if other forms of task-based supervision and task-based evaluation (Jing et al., 1998) may be better suited to the specialized task of automatic concept map summarization. For instance, instead of asking a human to identify better and worse summary statements, we might examine how well a user (or model) presented with summary statement s can answer if other summary statements s' are true or false. If some s helps identify many other true s', then s is (potentially) a good summary. We look forward to examining this idea in future work, following recent studies of question-based evaluation for the summarization task (Eyal et al., 2019).

8 Acknowledgement

Thanks to Haw-Shiuan Chang, Tu Vu and Kalpesh Krishna for helpful comments on earlier drafts of this work. Thanks to the anonymous reviewers for their helpful suggestions, in particular for pointing out possible connections between relationship summarization and joint extraction of relations and entities.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *NAACL*.

Tobias Falke and Iryna Gurevych. 2017. Bringing structure into summaries: Crowdsourcing a benchmark corpus of concept maps. In *EMNLP*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *NAACL*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL*.

Abram Handler, Matthew Denny, Hanna Wallach, and Brendan O'Connor. 2016. Bag of what? simple noun phrase extraction for text analysis. In *Proceedings of the First Workshop on NLP and Computational Social Science*.

Abram Handler and Brendan O'Connor. 2018. Relational summarization for corpus analysis. In *NAACL*.

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

Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1998. Summarization evaluation methods : Experiments and analysis. In *AAAI Spring Symposium*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *ACL*.

J.J. Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, University of Alberta.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *Proceedings of the eighth international conference on computational semantics*.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and hb. Technical report, Sawtooth Software.

Fabian Pedregosa-Izquierdo. 2015. *Feature extraction and supervised learning on fMRI : from practice to theory*. Theses, Université Pierre et Marie Curie - Paris VI.

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

Philip A Schrodt, Deborah J Gerner, Rajaa Abu-Jabr, Oemeur Yilmaz, and Erin M Simpson. 2001. Analyzing the dynamics of international mediation processes in the middle east and balkans. In *Annual Meeting of the American Political Science Association*.

John Sprouse and Carson Schütze. 2014. *Research Methods in Linguistics*, chapter Judgment Data. Cambridge University Press, Cambridge, UK.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *CoRR*, abs/1805.12471v1.