Extending the Entity-grid Coherence Model to Semantically Related Entities

Katja Filippova and **Michael Strube**
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/nlp/

Abstract

This paper reports on work in progress on extending the entity-based approach on measuring coherence (Barzilay & Lapata, 2005; Lapata & Barzilay, 2005) from coreference to semantic relatedness. We use a corpus of manually annotated German newspaper text (TüBa-D/Z) and aim at improving the performance by grouping related entities with the WikiRelate! API (Strube & Ponzetto, 2006).

1 Introduction

Evaluation is a well-known problem for Natural Language Generation (NLG). Human labor required to evaluate the output of a NLG system is expensive since every text should be read by several human judges and evaluated according to several parameters. Automatic summarization is an application using a NLG component which is hard to evaluate. The Document Understanding Conference[1], which every year issues a summarization task, distinguishes five aspects of linguistic quality of a summary: grammaticality, non-redundancy, referential clarity, focus and coherence. The parameter for which most participants get very low scores is coherence. This may reflect the difficulty which (mostly) extractive methods face during the ordering phase. Even if selected sentences are relevant and related, being in a wrong order they will make the summary hard to understand. The same is true for any other text-to-text generation system with a multisentential output.

In this paper we consider a way of automatic coherence assessment (Barzilay & Lapata, 2005) which is beneficial for such NLG systems. This method is based on how patterns of entity distribution differ for coherent and incoherent texts. It utilizes information of three kinds: coreference, salience and syntax. As a suggestion for future work, Barzilay & Lapata hypothesize that integrating semantic knowledge for entity grouping (as opposed to coreference) should improve the results. So, the purpose of the current study is threefold:

- to check how the method performs on a language other than English;
- to estimate the contribution of the three knowledge sources on mannualy annotated data;
- to see whether semantic clustering of entities outperfoms the coreference baseline.

2 The Entity-based Approach

Barzilay & Lapata (2005) describe a method for coherence assessment which grounds on the premises that (1) for a text to be globally coherent it has to be locally coherent as well; and (2) the patterns of how entities appear throughout the text differ for coherent and incoherent data.

To test their method, they consider a collection of coherent texts[2] and for each of them generate a number of incoherent variants by putting the sentences in a random order. Then, for each rendering, they create an *entity-grid* where each column represents an entity and each row represents a sentence from the text. A cell in a grid tells which syntactic function a given entity has in a given sentence. The set of possible functions is reduced to four: subject (**s**), object (**o**), other (**x**), or nothing (**-**) if the entity is not mentioned in a sentence. Two example grids

[1] http://duc.nist.gov

[2] They experiment with a corpus of newspaper articles and a corpus of accident reports, all in English.

	e_1	e_2	e_3	e_4	e_5	e_6
s_1	**S**	**O**	**X**	**-**	**-**	**-**
s_2	**O**	**S**	**-**	**O**	**-**	**-**
s_3	**S**	**-**	**-**	**-**	**-**	**-**
s_4	**-**	**-**	**-**	**-**	**-**	**S**

Table 1: Coherent text grid

	e_1	e_2	e_3	e_4	e_5	e_6
s_4	**-**	**-**	**-**	**-**	**-**	**S**
s_1	**S**	**O**	**X**	**-**	**-**	**-**
s_3	**S**	**-**	**-**	**-**	**-**	**-**
s_2	**O**	**S**	**-**	**O**	**-**	**-**

Table 2: Incoherent text grid

– for a coherent text and for its shuffled version – are presented in Tables 1 and 2 respectively.

To compare two texts which differ only in their sentence order, each of them is represented by a feature vector. A feature stands for a possible transition between syntactic functions of an entity (e.g. **-O**, **SX**, **SSO**). Unigram, bigram and trigram transitions are distinguished. The value of a transition feature is its probability calculated from the grid. For binary transitions there are, thus, 4×4 possible features. If there are no full parses available so that one cannot distinguish between syntactic realizations and fills a cell with **X** or **-** only, the number of binary transitions is reduced to $2 \times 2 = 4$. These simplified (i.e. without syntactic information) feature vectors for the grids in Tables 1 and 2 are given in Table 3.

	XX	**X-**	**-X**	**--**
g_1	0.17	0.28	0.17	0.39
g_2	0.11	0.22	0.33	0.33

Table 3: Feature vectors for grids in Tables 1 & 2

The coherence assessment is then formulated as a ranking learning problem. SVM^{light} (Joachims, 2002) is used for this task. Pairwise rankings (a coherent text vs. an incoherent rendering) are supplied to the learner as the relative quality of incoherent renderings is not known. For each document 20 pairs are generated in total.

Barzilay & Lapata (2005) obtain impressive results – about 90% of *ranking accuracy* which is the ratio of how often a coherent order is ranked higher than its incoherent variant[3]:

$$RA = \frac{correct\_pairs}{all\_pairs}$$

Barzilay & Lapata (2005) demonstrate that richer syntactic representation, as well as coreference resolution instead of string identity for entities identification, improve the performance. Another finding is that it is effective to distinguish between *salient* entities (those mentioned more than once: e_1, e_2 in Tables 1 & 2) and the rest. Given that they preprocess the data automatically by employing a state-of-the-art parser and a noun phrase coreference resolution system, manual annotation is expected to refine the model.

3 Reimplementation

We reimplemented the algorithm of Barzilay & Lapata (2005) and tested it on a German corpus of newpaper articles TüBa-D/Z (Telljohann et al., 2003). This corpus provides manual syntactic[4], morphological and NP coreference annotation (Hinrichs et al., 2004). We used the same SVM^{light} package for learning of a ranking function. Like Barzilay & Lapata, we took 100 articles for training, testing and development sets each. The results we report below are all computed from the development set. As results might differ considerably depending on how incoherent random orders are, for every article we continued to use the set of random orders generated during the first try. This allowed us to make objective judgements about the impact of a certain parameter on the performance. We also selected a subset of articles from the TüBa-D/Z in order to make the average article length equal to the average length of the articles Barzilay & Lapata used (i.e. 10.5 sentences).

3.1 Settings

Similar to Barzilay & Lapata, we experimented with the following settings:

COREF: coreference vs. word identity for entity identification;

SYNT: syntax-rich vs. simplified representation;

SAL: distinguishing between salient entities (mentioned exactly once) and the rest vs. without this distinction.

[3]Note, that random baseline ensures RA of 50%.

[4]Yannick Versley kindly helped us to to convert the syntactic annotation (Versley, 2005).

	+COREF	-COREF
+SYNT+SAL	72%	62%
+SYNT-SAL	69%	53%
-SYNT+SAL	75%	66%
-SYNT-SAL	71%	59%

Table 4: Ranking accuracy for different settings

The results for each of the settings are presented in Table 4. Although obtained from human-annotated data, they are strikingly lower than the results Barzilay & Lapata report for English. We concluded the following:

1. Coreference information definitely improves the performance. Using word match for entity clustering works only if combined with salience, otherwise the method is hardly better than the baseline.
2. The fact that quite some correct decisions could be made with all parameters set negative (-SYNT-SAL-COREF) brought us to the idea that there is a difference in the amount of entities mentioned in the first, the last and a middle sentences of a text. Having calculated the average number of entities[5] in these three types of sentences, we concluded that indeed the amount decreases as the text continues. In a coherent text the first sentence generally introduces more entities than any further sentence mentions. The last sentence is shorter and, on average, contains less entities than other ones.
3. Surprisingly, for our data syntactic information turned out to have a negative impact on the results, although it may be that a larger training set is needed to benefit from it.
4. The RA of 59% for -SYNT-SAL-COREF demonstrates that the method can be of use even if applied to data without any information but sentence boundaries.

3.2 Extended Rankings

Apart from the settings described above, we also experimented with the training data representation by extending the pairwise ranking to longer rankings. According to Lapata's (2006) psycholinguistic experiment, Kendall's τ correlates reliably with human judgements regarding ordering tasks. It varies between -1 and 1 and is calculated as $1 - 4\frac{t}{N(N-1)}$, where t is the number of interchanges of adjacent elements required to bring the total of N elements in the right order. Assuming that the lower the τ, the less coherent a text is, we supplied the learner with rankings of 3 sentences instead of pairwise rankings as well as with rankings of all 21 renderings. Unfortunately, this modification did not improve the results but caused a slight drop in performance: for the best setting (-SYNT+SAL+COREF) the RA was 73%.

[5] Both new and already mentioned entities count.

3.3 Beyond Entities

For entity clustering we used the WikiRelate! API (Strube & Ponzetto, 2006) to compute relatedness between entities. We preferred it to the GermaNet API (Gurevych, 2005) because the latter works better for computing semantic similarity whereas the former is more suitable for computing semantic relatedness. Apart from that, given that our data is a collection of newspaper articles containing named entities (persons, locations, organizations) which can be related as well, Wikipedia is a better choice as it covers named entities as well as common nouns (the version from 09/25/2006 has 471,065 entries). Future work should make use of both semantic resources. From the 6 possible measures implemented in WikiRelate!, we selected the Wu&Palmer measure as Strube & Ponzetto (2006) report that it demonstrated the highest correlation with humans.

The experiments with semantic relatedness had two goals:

- to see whether it can improve the best results achieved with coreference sets,
- to check whether semantic relatedness alone can be reliably used for entity clustering in case there is no coreference resolution system available.

Since syntactic information affects the results negatively while distinguishing between salient entites and the rest has a positive impact on them, we did not further experiment with all possible settings combinations and used -SYNT+SAL.

To group similar entities together, our algorithm proceeds as follows: when a new entity e_i is encountered, it is measured whether it is related to already found entities E. If there is an entity $e_j \in E$ such that $SemRel(e_i, e_j) > t$, where t is a threshold, then its history is further assigned to this entity. We experimented with different values for t:

the smaller the value, the denser the grid but the less related words within one entity group are.

t	-SYNT+SAL+COREF	-SYNT+SAL-COREF
w/o	75%	66%
0.1	71%	66%
0.2	72%	66%
0.3	72%	68%
0.4	73%	68%
0.5	73%	69%

Table 5: Ranking accuracy with different relatedness thresholds

The results demonstrate a significant improvement over the word-identity model although semantic relatedness is not as good as coreference, the difference between them still being about 5%. Semantic clustering of entities on top of coreference grouping does not bring an improvement, at least when done incrementally. A better approach might be to require any two entities from one cluster to have the minimum relatedness of t rather than adding an entity to a cluster when it is related to at least one element from the cluster.

4 Conclusions and Future Work

We presented our work on extending the entity-based coherence assessment from coreference to semantic relatedness and its application to German. In spite of the fact that we used human-annotated data, our results are considerably worse than the results for English. This may be caused by differences between the corpora. We analysed the impact of different settings and problem formulations (pairwise vs. multi-element rankings) and reported the best parameters for German.

Our initial experiments with entity clustering using semantic relatedness gave us some evidence that this is a promising direction to pursue. In particular, we would like to depart from the manually annotated data and explore cheaper approaches which require neither a deep parser, nor a coreference resolution system and work fully automatically. The RA of 69% obtained without syntactic and coreference information motivates this direction of research. Such an approach would provide a low-cost coherence evaluation strategy for NLG applications with a multisentential output.

Future work should compare (or combine) the information from Wikipedia with information from GermaNet and determine constraints on entity grouping. We experimented with incremental clustering although it may be that "shrinking" of a complete grid with a constraint on the size of a cluster would be more effective. We would also like to test the extended model on the data sets used by Barzilay & Lapata (2005).

Acknowledgements: This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a KTF grant (09.009.2004).

References

Barzilay, Regina & Mirella Lapata (2005). Modelling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, 25–30 June 2005, pp. 141–148.

Gurevych, Iryna (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, Jeju Island, South Korea, 11-13 October, 2005, pp. 767–778.

Hinrichs, Erhard, Sandra Kübler, Karin Naumann, Heike Telljohann & Julia Trushkina (2004). Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories*, Tübingen, Germany, 10–11 December 2004.

Joachims, Thorsten (2002). Optimizing search engines using clickthrough data. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 23–26 July 2002, pp. 133–142.

Lapata, Mirella (2006). Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

Lapata, Mirella & Regina Barzilay (2005). Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Edinburgh, Schotland, 30 July–5 August, 2005, pp. 1085–1090.

Strube, Michael & Simone Paolo Ponzetto (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, Mass., 16–20 July 2006, pp. 1219–1224.

Telljohann, Heike, Erhard Hinrichs & Sandra Kübler (2003). *Stylebook for the Tübingen treebank of written German (TüBa-D/Z)*. Technical Report: Seminar für Sprachwissenschaft, Universität Tübingen.

Versley, Yannick (2005). Parser evaluation across text types. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories*, Barcelona, Spain, 9-10 December 2005.