Domain-Independent Shallow Sentence Ordering

Thade Nahnsen
School of Informatics
University of Edinburgh
T.Nahnsen@sms.ed.ac.uk

Abstract

We present a shallow approach to the sentence ordering problem. The employed features are based on discourse entities, shallow syntactic analysis, and temporal precedence relations retrieved from VerbOcean. We show that these relatively simple features perform well in a machine learning algorithm on datasets containing sequences of events, and that the resulting models achieve optimal performance with small amounts of training data. The model does not yet perform well on datasets describing the consequences of events, such as the destructions after an earthquake.

1 Introduction

Sentence ordering is a problem in many natural language processing tasks. While it has, historically, mainly been considered a challenging problem in (concept-to-text) language generation tasks, more recently, the issue has also generated interest within summarization research (Barzilay, 2003; Ji and Pulman, 2006). In the spirit of the latter, this paper investigates the following questions: (1) Does the topic of the text influence the factors that are important to sentence ordering? (2) Which factors are most important for determining coherent sentence orderings? (3) How much performance is gained when using deeper knowledge resources?

Past research has investigated a wide range of aspects pertaining to the ordering of sentences in text. The most prominent approaches include: (1) temporal ordering in terms of publication date (Barzilay, 2003), (2) temporal ordering in terms of textual cues in sentences (Bollegala et al., 2006), (3) the topic of the sentences (Barzilay, 2003), (4) coherence theories (Barzilay and Lapata, 2008), e.g., Centering Theory, (5) content models (Barzilay and Lee, 2004), and (6) ordering(s) in the underlying documents in the case of summarisation (Bollegala et al., 2006; Barzilay, 2003).

2 The Model

We view coherence assessment, which we recast as a sentence ordering problem, as a machine learning problem using the feature representation discussed in Section 2.1. It can be viewed as a ranking task because a text can only be more or less coherent than some other text. The sentence ordering task used in this paper can easily be transformed into a ranking problem. Hence, paralleling Barzilay and Lapata (2008), our model has the following structure.

The data consists of alternative orderings (x_{ij}, x_{ik}) of the sentences of the same document d_i. In the training data, the preference ranking of the alternative orderings is known. As a result, training consists of determining a parameter vector $\mathbf{w}$ that minimizes the number of violations of pairwise rankings in the training set, a problem which can be solved using SVM constraint optimization (Joachims, 2002). The following section explores the features available for this optimization.

2.1 Features

Approaches to sentence ordering can generally be categorized as knowledge-rich or knowledge-lean. Knowledge-rich approaches rely on manually created representations of sentence orderings using do-

Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium, pages 78–83,
Boulder, Colorado, June 2009.

main communication knowledge.

Barzilay and Lee (2004)'s knowledge-lean approach attempts to automate the inference of knowledge-rich information using a distributional view of content. In essence, they infer a number of topics using clustering. The clusters are represented by corresponding states in a hidden Markov model, which is used to model the transitions between topics.

Lapata (2003), in contrast, does not attempt to model topics explicitly. Instead, she reduces sentence ordering to the task of predicting the next sentence given the previous sentence, which represents a coarse attempt at capturing local coherence constraints. The features she uses are derived from three categories - verbs, nouns, and dependencies - all of which are lexicalised. Her system thereby, to some extent, learns a precedence between the words in the sentences, which in turn represent topics.

Ji and Pulman (2006) base their ordering strategy not only on the directly preceding sentence, but on all preceding sentences. In this way, they are able to avoid a possible topic bias when summarizing multiple documents. This is specific to their approach as both Lapata (2003)'s and Barzilay and Lee (2004)'s approaches are not tailored to summarization and therefore do not experience the topic bias problem.

The present paper deviates from Lapata (2003) insofar as we do not attempt to learn the ordering preferences between pairs of sentences. Instead, we learn the ranking of documents. The advantage of this approach is that it allows us to straightforwardly discern the individual value of various features (cf. Barzilay and Lapata (2008)).

The methods used in this paper are mostly shallow with the exception of two aspects. First, some of the measures make use of WordNet relations (Fellbaum, 1998), and second, some use the temporal ordering provided by the "happens-before" relation in VerbOcean (Chklovski and Pantel, 2004). While the use of WordNet is self-explanatory, its effect on sentence ordering algorithms does not seem to have been explored in any depth. The use of VerbOcean is meant to reveal the degree to which common sense orderings of events affect the ordering of sentences, or whether the order is reversed.

With this background, the sentence ordering features used in this paper can be grouped into three categories:

2.1.1 Group Similarity

The features in this category are inspired by discourse entity-based accounts of local coherence. Yet, in contrast to Barzilay and Lapata (2008), who employ the syntactic properties of the respective occurrences, we reduce the accounts to whether or not the entities occur in subsequent sentences (similar to Karamanis (2004)'s NOCB metric). We also investigate whether using only the information from the head of the noun group (cf. Barzilay and Lapata (2008)) suffices, or whether performance is gained when allowing the whole noun group in order to determine similarity. Moreover, as indicated above, some of the noun group measures make use of WordNet synonym, hypernym, hyponym, antonym relationships. For completeness, we also consider the effects of using verb groups and whole sentences as syntactic units of choice.

2.1.2 Temporal Ordering

This set of features uses information on the temporal ordering of sentences, although it currently only includes the "happens-before" relations in VerbOcean.

2.1.3 Longer Range Relations

The group similarity features only capture the relation between a sentence and its immediate successor. However, the coherence of a text is clearly not only defined by direct relations, but also requires longer range relations between sentences (e.g., Barzilay and Lapata (2008)). The features in this section explore the impact of such relations on the coherence of the overall document as well as the appropriate way of modeling them.

3 Experiments

This section introduces the datasets used for the experiments, describes the experiments, and discusses our main findings.

3.1 Evaluation Datasets

The three datasets used for the automatic evaluation in this paper are based on human-generated texts (Table 1). The first two are the earthquake and accident datasets used by Barzilay and Lapata (2008).

Each of these sets consists of 100 datasets in the training and test sets, respectively, as well as 20 random permutations for each text.

The third dataset is similar to the first two in that it contains original texts and random permutations. In contrast to the other two sources, however, this dataset is based on the human summaries from DUC 2005 (Dang, 2005). It comprises 300 human summaries on 50 document sets, resulting in a total of 6,000 pairwise rankings split into training and test sets. The source furthermore differs from Barzilay and Lapata (2008)'s datasets in that the content of each text is not based on one individual event (an earthquake or accident), but on more complex topics followed over a period of time (e.g., the espionage case between GM and VW along with the various actions taken to resolve it). Since the different document sets cover completely different topics the third dataset will mainly be used to evaluate the topic-independent properties of our model.

Dataset	*Training*	*Testing*
Earthquakes	1,896	2,056
Accidents	2,095	2,087
DUC2005	up to 3,300	2,700

Table 1: Number of pairwise rankings in the training and test sets for the three datasets

3.2 Experiment 1

In the first part of this experiment, we consider the problem of the granularity of the syntactic units to be used. That is, does it make a difference whether we use the words in the sentence, the words in the noun groups, the words in the verb groups, or the words in the respective heads of the groups to determine coherence? (The units are obtained by processing the documents using the LT-TTT2 tools (Grover and Tobin, 2006); the lemmatizer used by LT-TTT2 is morpha (Minnen and Pearce, 2000).) We also consider whether lemmatization is beneficial in each of the granularities.

The results - presented in Table 2 - indicate that considering only the heads of the verb and noun groups separately provides the best performance. In particular, the heads outperform the whole groups, and the heads separately also outperform noun and verb group heads together. As for the question of whether lemmatization provides better results, one needs to distinguish the case of noun and verb groups. For noun groups, lemmatization improves performance, which can mostly be attributed to singular and plural forms. In the case of verb groups, however, the lemmatized version yields worse results than the surface forms, a fact mainly explained by the tense and modality properties of verbs.

Syntactic Unit	*Processing*	*Accuracy*	
		Acc	Earth
sentence	surface form	52.27	14.21
	lemma	52.27	12.04
heads sentence	surface form	77.35	60.30
	lemma	73.18	61.67
noun group	surface form	80.14	59.84
	lemma	81.58	59.54
head NG	surface form	80.49	59.75
	lemma	81.65	59.12
verb group	surface form	71.57	68.14
	lemma	53.40	68.01
head VG	surface form	71.15	68.39
	lemma	53.76	67.85

Table 2: Performance with respect to the syntactic unit of processing of the training datasets. Accuracy is the fraction of correctly ranked pairs of documents over the total number of pairs. (?Heads sentence? is the heads of NGs and VGs.)

Given the appropriate unit of granularity, we can consider the impact of semantic relations between surface realizations on coherence. For these experiments we use the synonym, hypernym, hyponym, and antonym relations in WordNet. The rationale for the consideration of semantic relations lies in the fact that the frequent use of the same words is usually deemed bad writing style. One therefore tends to observe the use of semantically similar terms in neighboring sentences. The results of using semantic relations for coherence rating are provided in Table 3. Synonym detection improves performance, while the other units provide poorer performance. This suggests that the hypernym and hyponym relations tend to over-generalize in the semantics.

The third category of features investigated is the temporal ordering of sentences; we use VerbOcean to obtain the temporal precedence between two events. One would expect events to be described ei-

Syntactic Unit	*Processing*	*Accuracy*	
		Acc	Earth
head NG	synonyms	82.37	59.40
	hypernyms	76.98	61.02
	hyponyms	81.59	59.14
	antonyms	74.20	48.07
	combines	70.84	56.51
head VG	synonyms	54.19	70.80
	hypernyms	53.36	60.54
	hyponyms	55.27	68.32
	antonyms	47.45	63.91
	combines	49.73	66.77

Table 3: The impact of WordNet on sentence ordering accuracy

Temporal Ordering	*Accuracy*	
	Acc	Earth
Precedence Ordering	60.41	47.09
Reverse Ordering	39.59	52.61
Precedence w/ matching NG	62.65	57.52
Reverse w/ matching NG	37.35	42.48

Table 4: The impact of the VerbOcean ?happens-before? temporal precedence relation on accuracy on the training datasets

ther in chronological order or in its reverse. While the former ordering represents a factual account of some sequence of events, the latter corresponds to newswire-style texts, which present the most important event(s) first, even though they may derive from previous events.

Table 4 provides the results of the experiments with temporal orderings. The first two rows validate the ordering of the events, while the latter two require the corresponding sentences to have a noun group in common in order to increase the likelihood that two events are related. The results clearly show that there is potential in the direct ordering of events. This suggests that sentence ordering can to some degree be achieved using simple temporal precedence orderings in a domain-independent way. This holds despite the results indicating that the features work better for sequences of events (as in the accident dataset) as opposed to accounts of the results of some event(s) (as in the earthquake dataset).

Range	*Accuracy*	
	Acc	Earth
2 occ. in 2 sent.	80.57	50.11
2 occ. in 3 sent.	73.17	45.43
3 occ. in 3 sent.	71.35	52.81
2 occ. in 4 sent.	66.95	50.41
3 occ. in 4 sent.	69.38	41.61
4 occ. in 4 sent.	71.93	58.97
2 occ. in 5 sent.	61.48	66.25
3 occ. in 5 sent.	68.59	42.33
4 occ. in 5 sent.	65.77	40.75
5 occ. in 5 sent.	81.39	62.40
sim. w/ sent. 1 sent. away	83.39	71.94
sim. w/ sent. 2 sent. away	60.44	67.52
sim. w/ sent. 3 sent. away	52.28	54.65
sim. w/ sent. 4 sent. away	49.65	44.50
sim. w/ sent. 5 sent. away	43.68	52.11

Table 5: Effect of longer range relations on coherence accuracy

The final category of features investigates the degree to which relations between sentences other than directly subsequent sentences are relevant. To this end, we explore two different approaches. The first set of features considers the distribution of entities within a fixed set of sentences, and captures in how many different sentences the entities occur. The resulting score is the number of times the entities occur in N out of M sentences. The second set only considers the similarity score from the current sentence and the other sentences within a certain range from the current sentence. The score of this feature is the sum of the individual similarities. Table 5 clearly confirms that longer range relations are relevant to the assessment of the coherence of text. An interesting difference between the two approaches is that sentence similarity only provides good results for neighboring sentences or sentences only one sentence apart, while the occurrence-counting method also works well over longer ranges.

Having evaluated the potential contributions of the individual features and their modeling, we now use SVMs to combine the features into one comprehensive measure. Given the indications from the foregoing experiments, the results in Table 6 are disappointing. In particular, the performance on the

Combination	*Accuracy*	
	Acc	Earth
Chunk+Temp+WN+LongRange+	83.11	54.88
Chunk+Temp+WN+LongRange-	77.67	62.76
Chunk+Temp+WN-LongRange+	74.17	59.28
Chunk+Temp+WN-LongRange-	68.15	63.55
Chunk+Temp-WN+LongRange+	*86.88*	*63.83*
Chunk+Temp-WN+LongRange-	80.19	59.43
Chunk+Temp-WN-LongRange+	76.63	60.86
Chunk+Temp-WN-LongRange-	64.43	60.94
NG Similarity w/ Synonyms	85.90	63.55
Coreference+Syntax+Salience+	90.4	87.2
Coreference-Syntax+Salience+	89.9	83.0
HMM-based Content Models	75.8	88.0
Latent Semantic Analysis	87.3	81.0

Table 6: Comparison of the developed model with other state-of-the-art systems. Coreference+Syntax+Salience+ and Coreference?Syntax+Salience+ are the Barzilay and Lapata (2008) model, HMM-based Content Models is the Barzilay and Lee (2004) paper and Latent Semantic Analysis is the Barzilay and Lapata (2008) implementation of Peter W. Foltz and Landauer (1998). The results of these systems are reproduced from Barzilay and Lapata (2008). (Temp = Temporal; WN = WordNet)

Coreference+Syntax+Salience+

Train \ Test	Earthquakes	Accidents
Earthquakes	87.3	67.0
Accidents	69.7	90.4

HMM-based Content Models

Train \ Test	Earthquakes	Accidents
Earthquakes	88.0	31.7
Accidents	60.3	75.8

Chunk+Temporal-WordNet+LongRange+

Train \ Test	Earthquakes	Accidents
Earthquakes	63.83	86.63
Accidents	64.19	86.88

Table 7: Cross-Training between Accident and Earthquake datasets. The results for Coreference+Syntax+Salience+ and HMM-Based Content Models are reproduced from Barzilay and Lapata (2008).

earthquake dataset is below standard. However, it seems that sentence ordering in that set is primarily defined by topics, as only content models perform well. (Barzilay and Lapata (2008) only perform well when using their coreference module, which determines antecedents based on the identified coreferences in the *original* sentence ordering, thereby biasing their orderings towards the correct ordering.) Longer range and WordNet relations together (Chunk+Temp-WN+LongRange+) achieve the best performance. The corresponding configuration is also the only one that achieves reasonable performance when compared with other systems.

4 Experiment 2

As stated, the ultimate goal of the models presented in this paper is the application of sentence ordering to automatically generated summaries. It is, in this regard, important to distinguish coherence as studied in Experiment 1 and coherence in the context of automatic summarization. Namely, for newswire summarization systems, the topics of the documents are unknown at the time of training. As a result, model performance on out-of-domain texts is important for summarization. Experiment 2 seeks to evaluate how well our model performs in such cases. To this end, we carry out two sets of tests. First, we cross-train the models between the accident and earthquake datasets to determine system performance in unseen domains. Second, we use the dataset based on the DUC 2005 model summaries to investigate whether our model's performance on unseen topics reaches a plateau after training on a particular number of different topics.

Surprisingly, the results are rather good, when compared to the poor results in part of the previous experiment (Table 7). In fact, model performance is nearly independent of the training topic. Nevertheless, the results on the earthquake test set indicate that our model is missing essential components for the correct prediction of sentence orderings on this set. When compared to the results obtained by Barzilay and Lapata (2008) and Barzilay and Lee (2004), it would appear that direct sentence-to-sentence similarity (as suggested by the Barzilay and Lapata baseline score) or capturing topic sequences are essential for acquiring the correct sequence of sentences in the earthquake dataset.

The final experimental setup applies the best

Different Topics	*Training Pairs*	*Accuracy*
2	160	55.17
4	420	63.54
6	680	65.20
8	840	65.57
10	1,100	64.80
15	1,500	64.93
20	2,100	64.87
25	2,700	64.94
30	3,300	65.61

Table 8: Accuracy on 20 test topics (2,700 pairs) with respect to the number of topics used for training using the model Chunk+Temporal-WordNet+LongRange+

model (Chunk+Temporal-WordNet+LongRange+) to the summarization dataset and evaluates how well the model generalises as the number of topics in the training dataset increases. The results - provided in Table 8 - indicate that very little training data (both regarding the number of pairs and the number of different topics) is needed. Unfortunately, they also suggest that the DUC summaries are more similar to the earthquake than to the accident dataset.

5 Conclusions

This paper investigated the effect of different features on sentence ordering. While a set of features has been identified that works well individually as well as in combination on the accident dataset, the results on the earthquake and DUC 2005 datasets are disappointing. Taking into account the performance of content models and the baseline of the Barzilay and Lapata (2008) model, the most convincing explanation is that the sentence ordering in the earthquake datasets is based on some sort of topic notion, providing a variety of possible antecedents between which our model is thus far unable to distinguish without resorting to the original (correct) ordering. Future work will have to concentrate on this aspect of sentence ordering, as it appears to coincide with the structure of the summaries for the DUC 2005 dataset.

References

Barzilay, R. (2003). *Information fusion for multi-document summarization: paraphrasing and generation*. Ph. D. thesis, Columbia University.

Barzilay, R. and M. Lapata (2008). Modeling local coherence: An entity-based approach. *Comput. Linguist. 34*, 1–34.

Barzilay, R. and L. Lee (2004). Catching the drift: probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL 2004*.

Bollegala, D., N. Okazaki, and M. Ishizuka (2006). A bottom-up approach to sentence ordering for multi-document summarization. In *Proceedings of ACL-44*.

Chklovski, T. and P. Pantel (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP 2004*.

Dang, H. (2005). Overview of duc 2005.

Fellbaum, C. (Ed.) (1998). *WordNet An Electronic Lexical Database*. The MIT Press.

Grover, C. and R. Tobin (2006). Rule-based chunking and reusability. In *Proceedings of LREC 2006*.

Ji, P. D. and S. Pulman (2006). Sentence ordering with manifold-based classification in multi-document summarization. In *Proceedings of EMNLP 2006*.

Joachims, T. (2002). Evaluating retrieval performance using clickthrough data. In *Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*.

Karamanis, N. (2004). Evaluating centering for sentence ordering in two new domains. In *Proceedings of the NAACL 2004*.

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proc. of ACL 2003*.

Minnen, G., C. J. and D. Pearce (2000). Robust, applied morphological generation. In *Proceedings of the 1st International Natural Language Generation Conference*.

Peter W. Foltz, W. K. and T. K. Landauer (1998). Textual coherence using latent semantic analysis. *Discourse Processes 25*, 285–307.