**Entity-based local coherence modelling using topological fields**

**Jackie Chi Kit Cheung and Gerald Penn** Department of Computer Science University of Toronto

Toronto, ON, M5S 3G4, Canada

*{*jcheung,gpenn*}*@cs.toronto.edu

**Abstract**

One goal of natural language generation is to produce coherent text that presents in- formation in a logical order. In this pa- per, we show that topological fields, which model high-level clausal structure, are an important component of local coherence in German. First, we show in a sen- tence ordering experiment that topologi- cal field information improves the entity grid model of Barzilay and Lapata (2008) more than grammatical role and simple clausal order information do, particularly when manual annotations of this informa- tion are not available. Then, we incor- porate the model enhanced with topolog- ical fields into a natural language gen- eration system that generates constituent orders for German text, and show that the added coherence component improves performance slightly, though not statisti- cally significantly.

**1 Introduction**

One type of coherence modelling that has captured recent research interest is local coherence mod- elling, which measures the coherence of a docu- ment by examining the similarity between neigh- bouring text spans. The entity-based approach, in particular, considers the occurrences of noun phrase entities in a document (Barzilay and Lap- ata, 2008). Local coherence modelling has been shown to be useful for tasks like natural language generation and summarization, (Barzilay and Lee,

2004) and genre classification (Barzilay and Lap- ata, 2008).

Previous work on English, a language with rel- atively fixed word order, has identified factors that contribute to local coherence, such as the gram- matical roles associated with the entities. There is

good reason to believe that the importance of these factors vary across languages. For instance, freer- word-order languages exhibit word order patterns which are dependent on discourse factors relating to information structure, in addition to the gram- matical roles of nominal arguments of the main verb. We thus expect word order information to be particularly important in these languages in dis- course analysis, which includes coherence mod- elling.

For example, Strube and Hahn (1999) introduce *Functional Centering*, a variant of Centering The- ory which utilizes information status distinctions between hearer-old and hearer-new entities. They apply their model to pronominal anaphora reso- lution, identifying potential antecedents of sub- sequent anaphora by considering syntactic and word order information, classifying constituents by their familiarity to the reader. They find that their approach correctly resolves more pronomi- nal anaphora than a grammatical role-based ap- proach which ignores word order, and the differ- ence between the two approaches is larger in Ger- man corpora than in English ones. Unfortunately, their criteria for ranking potential antecedents re- quire complex syntactic information in order to classify whether proper names are known to the hearer, which makes their algorithm hard to auto- mate. Indeed, all evaluation is done manually.

We instead use topological fields, a model of clausal structure which is indicative of information structure in German, but shallow enough to be au- tomatically parsed at high accuracy. We test the hypothesis that they would provide a good com- plement or alternative to grammatical roles in lo- cal coherence modelling. We show that they are superior to grammatical roles in a sentence or- dering experiment, and in fact outperforms sim- ple word-order information as well. We further show that these differences are particularly large when manual syntactic and grammatical role an-

186

*Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 186–195, Uppsala, Sweden, 11-16 July 2010. *Q*c 2010 Association for Computational Linguistics

S

VF LK MF

NF S

LK MF VC

Millionen von Mark verschwendet der Senat jeden Monat, weil er sparen will.

“The senate wastes millions of marks each month, because it wants to save.”

Figure 1: The clausal and topological field structure of a German sentence. Notice that the subordinate clause receives its own topology.

notations are not available.

We then embed these topological field annota- tions into a natural language generation system to show the utility of local coherence information in an applied setting. We add contextual features using topological field transitions to the model of Filippova and Strube (2007b) and achieve a slight improvement over their model in a con- stituent ordering task, though not statistically sig- nificantly. We conclude by discussing possible reasons for the utility of topological fields in lo- cal coherence modelling.

**2 Background and Related Work**

**2.1 German Topological Field Parsing**

Topological fields are sequences of one or more contiguous phrases found in an enclosing syntac- tic region, which is the clause in the case of the German topological field model (Ho¨ hle, 1983). These fields may have constraints on the number of words or phrases they contain, and do not nec- essarily form a semantically coherent constituent. In German, the topology serves to identify all of the components of the verbal head of a clause, as well as clause-level structure such as complemen- tizers and subordinating conjunctions. Topologi- cal fields are a useful abstraction of word order, because while Germanic word order is relatively free with respect to grammatical functions, the or- der of the topological fields is strict and unvarying.

A German clause can be considered to be an- chored by two “brackets” which contain modals, verbs and complementizers. The left bracket (*linke Klammer*, **LK**) may contain a complementizer, subordinating conjunction, or a finite verb, de- pending on the clause type, and the right bracket contains the verbal complex (**VC**). The other topo- logical fields are defined in relation to these two brackets, and contain all other parts of the clause

such as verbal arguments, adjuncts, and discourse cues.

The **VF** (*Vorfeld* or “pre-field”) is so-named be- cause it occurs before the left bracket. As the first constituent of most matrix clauses in declarative sentences, it has special significance for the coher- ence of a passage, which we will further discuss below. The **MF** (*Mittelfeld* or “middle field”) is the field bounded by the two brackets. Most verb arguments, adverbs, and prepositional phrases are found here, unless they have been fronted and put in the VF, or are prosodically heavy and postposed to the NF field. The **NF** (*Nachfeld* or “post-field”) contains prosodically heavy elements such as post- posed prepositional phrases or relative clauses, and occasionally postposed noun phrases.

**2.2 The Role of the Vorfeld**

One of the reasons that we use topological fields for local coherence modelling is the role that the VF plays in signalling the information structure of German clauses, as it often contains the topic of the sentence.

In fact, its role is much more complex than be- ing simply the topic position. Dipper and Zins- meister (2009) distinguish multiple uses of the VF depending on whether it contains an element re- lated to the surrounding discourse. They find that

45.1% of VFs are clearly related to the previous context by a reference or discourse relation, and a further 21.9% are deictic and refer to the situation described in the passage in a corpus study. They also run a sentence insertion experiment where subjects are asked to place an extracted sentence in its original location in a passage. The authors remark that extracted sentences with VFs that are referentially related to previous context (e.g., they contain a coreferential noun phrase or a discourse relation like “therefore”) are reinserted at higher accuracies.

a)

|  |  |
| --- | --- |
| **#** | **Original Sentence and Translation** |
| 1 | *Einen Zufluchtsort fu¨ r Frauen, die von ihren Ma¨ nnern mißhandelt werden, gibt es nunmehr auch*  *in Treptow.*  “There is now a sanctuary for women who are mistreated by their husbands in Treptow as well.” |
| 2 | *Das Bezirksamt bietet Frauen (auch mit Kindern) in derartigen Notsituationen voru¨ bergehend*  *eine Unterkunft.*  “The district office offers women (even with children) in this type of emergency temporary  accommodation.” |
| 3 | *Zugleich werden die Betroffenen der Regelung des Unterhalts, bei Beho¨ rdenga¨ ngen und auch*  *bei der Wohnungssuche unterstu¨ tzt.*  “At the same time, the affected are supported with provisions of necessities, in dealing with  authorities, and also in the search for new accommodations.” |

b)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *DE*  EN | *Zufluchtsort*  sanctuary | *Frauen*  women | *Ma¨ nnern*  husbands | *Treptow*  Treptow | *Kindern*  children |
| 1  2  3 | acc  *−*  *−* | oth  oth nom | oth  *−*  *−* | oth  *−*  *−* | *−*  oth  *−* |

c)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *− −* | *−* nom | *−* acc | *−* oth | nom *−* | nom nom | nom acc | nom oth |
| 0.3 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| acc *−* | acc nom | acc acc | acc oth | oth *−* | oth nom | oth acc | oth oth |
| 0.1 | 0.0 | 0.0 | 0.0 | 0.3 | 0.1 | 0.0 | 0.1 |

Table 1: a) An example of a document from Tu¨ Ba-D/Z, b) an abbreviated entity grid representation of it, and c) the feature vector representation of the abbreviated entity grid for transitions of length two. Mentions of the entity *Frauen* are underlined. nom: nominative, acc: accusative, oth: dative, oblique, and other arguments

Filippova and Strube (2007c) also examine the role of the VF in local coherence and natural lan- guage generation, focusing on the correlation be- tween VFs and sentential topics. They follow Ja- cobs (2001) in distinguishing the *topic of addres- sation*, which is the constituent for which the proposition holds, and *frame-setting topics*, which is the domain in which the proposition holds, such as a temporal expression. They show in a user study that frame-setting topics are preferred to top- ics of addressation in the VF, except when a con- stituent needs to be established as the topic of ad- dressation.

**2.3 Using Entity Grids to Model Local**

**Coherence**

Barzilay and Lapata (2008) introduce the entity grid as a method of representing the coherence of a document. Entity grids indicate the location of the occurrences of an entity in a document, which is

important for coherence modelling because men- tions of an entity tend to appear in clusters of neighbouring or nearby sentences in coherent doc- uments. This last assumption is adapted from Cen- tering Theory approaches to discourse modelling.

In Barzilay and Lapata (2008), an entity grid is constructed for each document, and is represented as a matrix in which each row represents a sen- tence, and each column represents an entity. Thus, a cell in the matrix contains information about an entity in a sentence. The cell is marked by the presence or absence of the entity, and can also be augmented with other information about the en- tity in this sentence, such as the grammatical role of the noun phrase representing that entity in that sentence, or the topological field in which the noun phrase appears.

Consider the document in Table 1. An entity grid representation which incorporates the syntac- tic role of the noun phrase in which the entity ap-

pears is also shown (not all entities are listed for brevity). We tabulate the transitions of entities be- tween different syntactic positions (or their non- occurrence) in sentences, and convert the frequen- cies of transitions into a feature vector representa- tion of transition probabilities in the document.

To calculate transition probabilities, we divide the frequency of a particular transition by the total number of transitions of that length.

This model of local coherence was investigated for German by Filippova and Strube (2007a). The main focus of that work, however, was to adapt the model for use in a low-resource situation when perfect coreference information is not available. This is particularly useful in natural language un- derstanding tasks. They employ a semantic clus- tering model to relate entities. In contrast, our work focuses on improving performance by anno- tating entities with additional linguistic informa- tion, such as topological fields, and is geared to- wards natural language generation systems where perfect information is available.

Similar models of local coherence include vari- ous Centering Theory accounts of local coherence ((Kibble and Power, 2004; Poesio et al., 2004) *inter alia*). The model of Elsner and Charniak (2007) uses syntactic cues to model the discourse- newness of noun phrases. There are also more global content models of topic shifts between sen- tences like Barzilay and Lee (2004).

**3 Sentence Ordering Experiments**

**3.1 Method**

We test a version of the entity grid representa- tion augmented with topological fields in a sen- tence ordering experiment corresponding to Ex- periment 1 of Barzilay and Lapata (2008). The task is a binary classification task to identify the original version of a document from another ver- sion which contains the sentences in a randomly permuted order, which is taken to be incoherent. We solve this problem in a supervised machine learning setting, where the input is the feature vec- tor representations of the two versions of the doc- ument, and the output is a binary value indicating the document with the original sentence ordering. We use SVMlight’s ranking module for classifi- cation (Joachims, 2002).

The corpus in our experiments consists of the last 480 documents of Tu¨ Ba-D/Z version 4 (Telljo- hann et al., 2004), which contains manual corefer-

ence, grammatical role and topological field infor- mation. This set is larger than the set that was used in Experiment 1 of Barzilay and Lapata (2008), which consists of 400 documents in two English subcorpora on earthquakes and accidents respec- tively. The average document length in the Tu¨ Ba- D/Z subcorpus is also greater, at 19.2 sentences compared to about 11 for the two subcorpora. Up to 20 random permutations of sentences were gen- erated from each document, with duplicates re- moved.

There are 216 documents and 4126 original- permutation pairs in the training set, and 24 docu- ments and 465 pairs in the development set. The remaining 240 documents are in the final test set (4243 pairs). The entity-based model is parame- terized as follows.

*Transition length* – the maximum length of the transitions used in the feature vector representa- tion of a document.

*Representation* – when marking the presence of an entity in a sentence, what information about the entity is marked (topological field, grammat- ical role, or none). We will describe the represen- tations that we try in section 3.2.

*Salience* – whether to set a threshold for the fre- quency of occurrence of entities. If this is set, all entities below a certain frequency are treated sep- arately from those reaching this frequency thresh- old when calculating transition probabilities. In the example in Table 1, with a salience thresh- old of 2, *Frauen* would be treated separately from *Ma¨ nnern* or *Kindern*.

Transition length, salience, and a regularization

parameter are tuned on the development set. We only report results using the setting of transition length *≤* 4, and no salience threshold, because they give the best performance on the development set. This is in contrast to the findings of Barzi- lay and Lapata (2008), who report that transition length *≤* 3 and a salience threshold of 2 perform best on their data.

**3.2 Entity Representations**

The main goal of this study is to compare word order, grammatical role and topological field in- formation, which is encoded into the entity grid at each occurrence of an entity. Here, we describe the variants of the entity representations that we compare.

**Baseline Representations** We implement sev- eral baseline representations against which we test our topological field-enhanced model. The sim- plest baseline representation marks the mere ap- pearance of an entity without any additional infor- mation, which we refer to as default.

Another class of baseline representations mark

the order in which entities appear in the clause. The correlation between word order and informa- tion structure is well known, and has formed the basis of some theories of syntax such as the Prague School’s (Sgall et al., 1986). The two versions of clausal order we tried are order 1/2/3+, which marks a noun phrase as the first, the sec- ond, or the third or later to appear in a clause, and order 1/2+, which marks a noun phrase as the first, or the second or later to appear in a clause. Since noun phrases can be embedded in other noun phrases, overlaps can occur. In this case, the dominating noun phrase takes the smallest order number among its dominated noun phrases.

The third class of baseline representations we employ mark an entity by its grammatical role in the clause. Barzilay and Lapata (2008) found that grammatical role improves performance in this task for an English corpus. Because Ger- man distinguishes more grammatical roles mor- phologically than English, we experiment with various granularities of role labelling. In particu- lar, subj/obj distinguishes the subject position, the object position, and another category for all other positions. cases distinguishes five types of entities corresponding to the four morphological cases of German in addition to another category for noun phrases which are not complements of the main verb.

**Topological Field-Based** These representations mark the topological field in which an entity ap- pears. Some versions mark entities which are prepositional objects separately. We try versions which distinguish VF from non-VF, as well as more general versions that make use of a greater set of topological fields. vf marks the noun phrase as belonging to a VF (and not in a PP) or not. vfpp is the same as above, but allows preposi- tional objects inside the VF to be marked as VF. topf/pp distinguishes entities in the topological fields VF, MF, and NF, contains a separate cat- egory for PP, and a category for all other noun phrases. topf distinguishes between VF, MF, and NF, on the one hand, and everything else on the

other. Prepositional objects are treated the same as other noun phrases here.

**Combined** We tried a representation which combines grammatical role and topological field into a single representation, subj/obj*×*vf, which takes the Cartesian product of subj/obj and vf above.

Topological fields do not map directly to topic-

focus distinctions. For example, besides the topic of the sentence, the Vorfeld may contain discourse cues, expletive pronouns, or the informational or contrastive focus. Furthermore, there are addi- tional constraints on constituent order related to pronominalization. Thus, we devised additional entity representations to account for these aspects of German.

topic attempts to identify the sentential topic of a clause. A noun phrase is marked as TOPIC if it is in VF as in vfpp, or if it is the first noun phrase in MF and also the first NP in the clause. Other noun phrases in MF are marked as NONTOPIC. Categories for NF and miscella- neous noun phrases also exist. While this repre- sentation may appear to be very similar to sim- ply distinguishing the first entity in a clause as for order 1/2+ in that TOPIC would correspond to the first entity in the clause, they are in fact dis- tinct. Due to issues related to coordination, appos- itive constructions, and fragments which do not receive a topology of fields, the first entity in a clause is labelled the TOPIC only 80.8% of the time in the corpus. This representation also distin- guishes NFs, which clausal order does not.

topic+pron refines the above by taking into account a word order restriction in German that pronouns appear before full noun phrases in the MF field. The following set of decisions repre- sents how a noun phrase is marked: If the first NP in the clause is a pronoun in an MF field and is the subject, we mark it as TOPIC. If it is not the sub- ject, we mark it as NONTOPIC. For other NPs, we follow the topic representation.

**3.3 Automatic annotations**

While it is reasonable to assume perfect annota- tions of topological fields and grammatical roles in many NLG contexts, this assumption may be less appropriate in other applications involving text-to- text generation where the input to the system is text such as paraphrasing or machine translation. Thus, we test the robustness of the entity repre-

|  |  |
| --- | --- |
| *Annotation* | *Accuracy* (%) |
| Grammatical role | 83.6 |
| Topological field (+PP) | 93.8 |
| Topological field (*−*PP) | 95.7 |
| Clausal order | 90.8 |

Table 2: Accuracy (%) of the permutation de- tection experiment with various entity represen- tations using manual and automatic annotations of topological fields and grammatical roles. The baseline without any additional annotation is un- derlined. Two-tailed sign tests were calculated for each result against the best performing model in each column (1: *p* = 0*.*101; 2: *p* = 0*.*053; +: statis- tically significant, *p <* 0*.*05; ++: very statistically significant, *p <* 0*.*01 ).

|  |  |
| --- | --- |
| *Representation* | *Manual Automatic* |
| topf/pp | **94.44 94.89** |
| topic | 94.13 94.53 |
| topic+pron | 94.08 94.51 |
| topf | 93.87 93.11 |
| subj/obj | 93.831 91.7++ |
| cases | 93.312 90.93++ |
| order 1/2+ | 92.51++ 92.1+ |
| subj/obj*×*vf | 92.32++ 90.74++ |
| default | 91.42++ 91.42++ |
| vfpp | 91.37++ 91.68++ |
| vf | 91.21++ 91.16++ |
| order 1/2/3+ | 91.16++ 90.71++ |

sentations to automatic extraction in the absence of manual annotations. We employ the following two systems for extracting topological fields and grammatical roles.

To parse topological fields, we use the Berke- ley parser of Petrov and Klein (2007), which has been shown to perform well at this task (Cheung and Penn, 2009). The parser is trained on sections of Tu¨ Ba-D/Z which do not overlap with the sec- tion from which the documents for this experiment were drawn, and obtains an overall parsing per- formance of 93.35% *F*1 on topological fields and clausal nodes without gold POS tags on the section of Tu¨ Ba-D/Z it was tested on.

We tried two methods to obtain grammatical roles. First, we tried extracting grammatical roles from the parse trees which we obtained from the Berkeley parser, as this information is present in the edge labels that can be recovered from the parse. However, we found that we achieved bet- ter accuracy by using RFTagger (Schmid and Laws, 2008), which tags nouns with their morpho- logical case. Morphological case is distinct from grammatical role, as noun phrases can function as adjuncts in possessive constructions and preposi-

Table 3: Accuracy of automatic annotations of noun phrases with coreferents. +PP means that prepositional objects are treated as a separate cate- gory from topological fields. *−*PP means they are treated as other noun phrases.

tional phrases. However, we can approximate the grammatical role of an entity using the morpho- logical case. We follow the annotation conven- tions of Tu¨ Ba-D/Z in not assigning a grammati- cal role when the noun phrase is a prepositional object. We also do not assign a grammatical role when the noun phrase is in the genitive case, as genitive objects are very rare in German and are far outnumbered by the possessive genitive con- struction.

**3.4 Results**

Table 2 shows the results of the sentence ordering permutation detection experiment. The top four performing entity representations are all topologi- cal field-based, and they outperform grammatical role-based and simple clausal order-based mod- els. These results indicate that the information that topological fields provide about clause struc- ture, appositives, right dislocation, etc. which is not captured by simple clausal order is important for coherence modelling. The representations in- corporating linguistics-based heuristics do not out- perform purely topological field-based models.

Surprisingly, the VF-based models fare quite poorly, performing worse than not adding any an- notations, despite the fact that topological field- based models in general perform well. This result may be a result of the heterogeneous uses of the VF.

The automatic topological field annotations are more accurate than the automatic grammatical role annotations (Table 3), which may partly explain why grammatical role-based models suffer more when using automatic annotations. Note, how- ever, that the models based on automatic topolog- ical field annotations outperform even the gram- matical role-based models using manual annota- tion (at marginal significance, *p <* 0*.*1). The topo-

logical field annotations are accurate enough that automatic annotations produce no decrease in per- formance.

These results show the upper bound of entity- based local coherence modelling with perfect coreference information. The results we obtain are higher than the results for the English cor- pora of Barzilay and Lapata (2008) (87.2% on the Earthquakes corpus and 90.4% on the Accidents corpus), but this is probably due to corpus differ- ences as well as the availability of perfect corefer- ence information in our experiments1.

Due to the high performance we obtained, we calculated Kendall tau coefficients (Lapata, 2006) over the sentence orderings of the cases in which our best performing model is incorrect, to deter- mine whether the remaining errors are instances where the permuted ordering is nearly identical to the original ordering. We obtained a *τ* of 0.0456 in these cases, compared to a *τ* of *−*0.0084 for all the pairs, indicating that this is not the case.

To facilitate comparison to the results of Filip- pova and Strube (2007a), we rerun this experiment on the same subsections of the corpus as in that work for training and testing. The first 100 arti- cles of Tu¨ Ba-D/Z are used for testing, while the next 200 are used for training and development.

Unlike the previous experiments, we do not do parameter tuning on this set of data. Instead, we follow Filippova and Strube (2007a) in using tran- sition lengths of up to three. We do not put in a salience threshold. We see that our results are much better than the ones reported in that work, even for the default representation. The main reason for this discrepancy is probably the way that entities are created from the corpus. In our experiments, we create an entity for every single noun phrase node that we encounter, then merge the entities that are linked by coreference. Filip- pova and Strube (2007a) convert the annotations of Tu¨ Ba-D/Z into a dependency format, then ex- tract entities from the noun phrases found there. They may thus annotate fewer entities, as there

1 Barzilay and Lapata (2008) use the coreference sys- tem of Ng and Cardie (2002) to obtain coreference anno- tations. We are not aware of similarly well-tested, pub- licly available coreference resolution systems that handle all types of anaphora for German. We considered adapting the BART coreference resolution toolkit (Versley et al., 2008) to German, but a number of language-dependent decisions re- garding preprocessing, feature engineering, and the learning paradigm would need to be made in order to achieve rea- sonable performance comparable to state-of-the-art English coreference resolution systems.

|  |  |
| --- | --- |
| *Representation* | *Accuracy* (%) |
| topf/pp | **93.83** |
| topic | 93.31 |
| topic+pron | 93.31 |
| topf | 92.49 |
| subj/obj | 88.99 |
| order 1/2+ | 88.89 |
| order 1/2/3+ | 88.84 |
| cases | 88.63 |
| vf | 87.60 |
| vfpp | 88.17 |
| default | 87.55 |
| subj/obj*×*vf | 87.71 |
| (Filippova and Strube, 2007) | 75 |

Table 4: Accuracy (%) of permutation detection experiment with various entity representations us- ing manual and automatic annotations of topolog- ical fields and grammatical roles on subset of cor- pus used by Filippova and Strube (2007a).

may be nested NP nodes in the original corpus. There may also be noise in the dependency con- version process.

The relative rankings of different entity repre- sentations in this experiment are similar to the rankings of the previous experiment, with topolog- ical field-based models outperforming grammati- cal role and clausal order models.

**4 Local Coherence for Natural Language**

**Generation**

One of the motivations of the entity grid-based model is to improve surface realization decisions in NLG systems. A typical experimental design would pass the contents of the test section of a corpus as input to the NLG system with the order- ing information stripped away. The task is then to regenerate the ordering of the information found in the original corpus. Various coherence models have been tested in corpus-based NLG settings. For example, Karamanis et al. (2009) compare several versions of Centering Theory-based met- rics of coherence on corpora by examining how highly the original ordering found in the corpus is ranked compared to other possible orderings of propositions. A metric performs well if it ranks the original ordering better than the alternative or- derings.

In our next experiment, we incorporate local co-

herence information into the system of Filippova and Strube (2007b). We embed entity topologi- cal field transitions into their probabilistic model, and show that the added coherence component slightly improves the performance of the baseline NLG system in generating constituent orderings in a German corpus, though not to a statistically sig- nificant degree.

**4.1 Method**

We use the WikiBiography corpus2 for our exper- iments. The corpus consists of more than 1100 bi- ographies taken from the German Wikipedia, and contains automatic annotations of morphological, syntactic, and semantic information. Each article also contains the coreference chain of the subject of the biography (the biographee). The first 100 articles are used for testing, the next 200 for de- velopment, and the rest for training.

The baseline generation system already incor- porates topological field information into the con- stituent ordering process. The system operates in two steps. First, in main clauses, one constituent is selected as the Vorfeld (VF). This is done us- ing a maximum entropy model (call it MAXENT). Then, the remaining constituents are ordered using a second maximum entropy model (MAXENT2). Significantly, Filippova and Strube (2007b) found that selecting the VF first, and then ordering the remaining constituents results in a 9% absolute improvement over the corresponding model where the selection is performed in one step by the sort- ing algorithm alone.

The maximum entropy model for both steps rely on the following features:

*•* features on the voice, valency, and identity of the main verb of the clause

*•* features on the morphological and syntactic status of the constituent to be ordered

*•* whether the constituent occurs in the preced- ing sentence

*•* features for whether the constituent contains a determiner, an anaphoric pronoun, or a rel- ative clause

*•* the size of the constituent in number of mod- ifiers, in depth, and in number of words

2 <http://www.eml-research.de/english/>

research/nlp/download/wikibiography.php

*•* the semantic class of the constituent (per- son, temporal, location, etc.) The biographee, in particular, is marked by its own semantic class.

In the first VF selection step, MAXENT simply produces a probability of each constituent being a VF, and the constituent with the highest probabil- ity is selected. In the second step, MAXENT2 takes the featural representation of two constituents, and produces an output probability of the first con- stituent preceding the second constituent. The fi- nal ordering is achieved by first randomizing the order of the constituents in a clause (besides the first one, which is selected to be the VF), then sorting them according to the precedence proba- bilities. Specifically, a constituent A is put before a constituent B if MAXENT2(A,B) *>* 0.5. Because this precedence relation is not antisymmetric (i.e., MAXENT2(A,B) *>* 0.5 and MAXENT2(B,A) *>*

0.5 may be simultaneously true or simultaneously

false), different initializations of the order pro- duce different sorted results. In our experiments, we correct this by defining the precedence rela- tion to be A precedes B iff MAXENT2(A,B) *>* MAXENT2(B,A). This change does not greatly im- pact the performance, and removes the random- ized element of the algorithm.

The baseline system does not directly model the context when ordering constituents. All of the features but one in the original maximum entropy models rely on local properties of the clause. We incorporate local coherence information into the model by adding entity transition features which we found to be useful in the sentence ordering ex- periment in Section 3 above.

Specifically, we add features indicating the topological fields in which entities occur in the previous sentences. We found that looking back up to two sentences produces the best results (by tuning on the development set). Because this cor- pus does not come with general coreference in- formation except for the coreference chain of the biographee, we use the semantic classes instead. So, all constituents in the same semantic class are treated as one coreference chain. An example of a feature may be **biog-last2**, which takes on a value such as ‘v*−*’, meaning that this constituent refers to the biographee, and the biographee occurs in the VF two clauses ago (v), but does not appear in the previous clause (*−*). For a constituent which is not the biographee, this feature would be marked

|  |  |
| --- | --- |
| *Method* | *VF Acc* (%) *Acc* (%) *Tau* |
| Baseline | 68.7 60.9 **0.72** |
| +Coherence | **69.2 61.5 0.72** |

Table 5: Results of adding coherence features into a natural language generation system. VF Acc% is the accuracy of selecting the first constituent in main clauses. Acc % is the percentage of per- fectly ordered clauses, tau is Kendall’s *τ* on the constituent ordering. The test set contains 2246 clauses, of which 1662 are main clauses.

‘na’ (not applicable).

**4.2 Results**

Table 5 shows the results of adding these contex- tual features into the maximum entropy models. We see that we obtain a small improvement in the accuracy of VF selection, and in the accuracy of correctly ordering the entire clause. These im- provements are not statistically significant by Mc- Nemar’s test. We suggest that the lack of coref- erence information for all entities in the article may have reduced the benefit of the coherence component. Also, the topline of performance is substantially lower than 100%, as multiple order- ings are possible and equally valid. Human judge- ments on information structuring for both inter- and intra-sentential units are known to have low agreement (Barzilay et al., 2002; Filippova and Strube, 2007c; Lapata, 2003; Chen et al., 2007). Thus, the relative error reduction is higher than the absolute reduction might suggest.

**5 Conclusions**

We have shown that topological fields are a use- ful source of information for local coherence mod- elling. In a sentence-order permutation detection task, models which use topological field infor- mation outperform both grammatical role-based models and models based on simple clausal or- der, with the best performing model achieving a relative error reduction of 40.4% over the original baseline without any additional annotation. Ap- plying our local coherence model in another set- ting, we have embedded topological field transi- tions of entities into an NLG system which orders constituents in German clauses. We find that the coherence-enhanced model slightly outperforms the baseline system, but this was not statistically significant.

We suggest that the utility of topological fields in local coherence modelling comes from the in- teraction between word order and information structure in freer-word-order languages. Crucially, topological fields take into account issues such as coordination, appositives, sentential fragments and differences in clause types, which word or- der alone does not. They are also shallow enough to be accurately parsed automatically for use in resource-poor applications.

Further refinement of the topological field an- notations to take advantage of the fact that they do not correspond neatly to any single information status such as topic or focus could provide addi- tional performance gains. The model also shows promise for other discourse-related tasks such as coreference resolution and discourse parsing.

**Acknowledgements**

We are grateful to Katja Filippova for providing us with source code for the experiments in Section 4 and for answering related questions, and to Tim- othy Fowler for useful discussions and comments on a draft of the paper. This work is supported in part by the Natural Sciences and Engineering Re- search Council of Canada.

**References**

R. Barzilay and M. Lapata. 2008. Modeling local co- herence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

R. Barzilay and L. Lee. 2004. Catching the drift: Prob- abilistic content models, with applications to gen- eration and summarization. In *Proc. HLT-NAACL*

*2004*, pages 113–120.

R. Barzilay, N. Elhadad, and K. McKeown. 2002. In- ferring strategies for sentence ordering in multidoc- ument news summarization. *Journal of Artificial In- telligence Research*, 17:35–55.

E. Chen, B. Snyder, and R. Barzilay. 2007. Incremen- tal text structuring with online hierarchical ranking. In *Proceedings of EMNLP*, pages 83–91.

J.C.K. Cheung and G. Penn. 2009. Topological Field Parsing of German. In *Proc. 47th ACL and 4th IJC- NLP*, pages 64–72. Association for Computational Linguistics.

S. Dipper and H. Zinsmeister. 2009. The Role of the German Vorfeld for Local Coherence: A Pi- lot Study. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 69–79. Gunter Narr.

M. Elsner and E. Charniak. 2007. A generative discourse-new model for text coherence. Technical report, Technical Report CS-07-04, Brown Univer- sity.

K. Filippova and M. Strube. 2007a. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages

139–142. Association for Computational Linguis- tics.

K. Filippova and M. Strube. 2007b. Generating con- stituent order in German clauses. In *Proc. 45th ACL*, pages 320–327.

K. Filippova and M. Strube. 2007c. The German Vor- feld and Local Coherence. *Journal of Logic, Lan- guage and Information*, 16(4):465–485.

T.N. Ho¨ hle. 1983. *Topologische Felder*. Ph.D. thesis, Ko¨ ln.

J. Jacobs. 2001. The dimensions of topiccomment.

*Linguistics*, 39(4):641–681.

T. Joachims. 2002. *Learning to Classify Text Using*

*Support Vector Machines*. Kluwer.

N. Karamanis, C. Mellish, M. Poesio, and J. Oberlan- der. 2009. Evaluating centering for information or- dering using corpora. *Computational Linguistics*,

35(1):29–46.

R. Kibble and R. Power. 2004. Optimizing referential coherence in text generation. *Computational Lin- guistics*, 30(4):401–416.

M. Lapata. 2003. Probabilistic text structuring: Exper- iments with sentence ordering. In *Proc. 41st ACL*, pages 545–552.

M. Lapata. 2006. Automatic evaluation of information ordering: Kendall’s tau. *Computational Linguistics*,

32(4):471–484.

V. Ng and C. Cardie. 2002. Improving machine learn- ing approaches to coreference resolution. In *Proc.*

*40th ACL*, pages 104–111.

S. Petrov and D. Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411.

M. Poesio, R. Stevenson, B.D. Eugenio, and J. Hitze- man. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*,

30(3):309–363.

H. Schmid and F. Laws. 2008. Estimation of condi- tional probabilities with decision trees and an appli- cation to fine-grained POS tagging. In *Proc. 22nd COLING*, pages 777–784. Association for Compu- tational Linguistics.

P. Sgall, E. Hajicˇova´, J. Panevova´, and J. Mey. 1986.

*The meaning of the sentence in its semantic and pragmatic aspects*. Springer.

M. Strube and U. Hahn. 1999. Functional center- ing: Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–

344.

H. Telljohann, E. Hinrichs, and S. Kubler. 2004.

The Tu¨ Ba-D/Z treebank: Annotating German with a context-free backbone. In *Proc. Fourth Interna- tional Conference on Language Resources and Eval- uation (LREC 2004)*, pages 2229–2235.

Y. Versley, S.P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, and A. Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Proc. 46th ACL-HLT Demo Session*, pages 9–12. Association for Computational Linguistics.