Three Knowledge-Free Methods for Automatic Lexical Chain Extraction

Steffen Remus and **Chris Biemann**
FG Language Technology
Department of Computer Science
Technische Universität Darmstadt
remus@kdsl.informatik.tu-darmstadt.de, biem@cs.tu-darmstadt.de

Abstract

We present three approaches to lexical chaining based on the LDA topic model and evaluate them intrinsically on a manually annotated set of German documents. After motivating the choice of statistical methods for lexical chaining with their adaptability to different languages and subject domains, we describe our new two-level chain annotation scheme, which rooted in the concept of cohesive harmony. Also, we propose a new measure for direct evaluation of lexical chains. Our three LDA-based approaches outperform two knowledge-based state-of-the art methods to lexical chaining by a large margin, which can be attributed to lacking coverage of the knowledge resource. Subsequent analysis shows that the three methods yield a different chaining behavior, which could be utilized in tasks that use lexical chaining as a component within NLP applications.

1 Introduction

A text that is understandable by its nature exhibits an underlying structure which makes the text coherent; that is, the structure is responsible for making the text "hang" together (Halliday and Hasan, 1976). The theoretic foundation of this structure is defined as *coherence* and *cohesion*. While the former is concerned with the meaning of a text, the latter can be seen as a collection of devices for creating it. Cohesion and coherence build the basis for most of the current natural language processing problems that deal with text understanding. *Lexical cohesion* ties together words or phrases that are semantically related. Once all the cohesive ties are identified the involved items can be grouped together to form so-called *lexical chains*, which form a theoretically well-founded building block in various natural language processing applications, such as word sense disambiguation (Okumura and Honda, 1994), summarization (Barzilay and Elhadad, 1997), malapropism detection and correction (Hirst and St-Onge, 1998), document hyperlinking (Green, 1996), text segmentation (Stokes et al., 2004), topic tracking (Carthy, 2004), and others. The performance of the individual task heavily depends on the quality of the identified lexical chains.

1.1 Motivation for Corpus-driven Approach

Previous approaches mainly focus on the use of knowledge resources like lexical semantic databases (Hirst and St-Onge, 1998) or thesauri (Morris and Hirst, 1991) as background information in order to resolve possible semantic relations. A major drawback of this strategy is the dependency on the coverage of the resource, which has a direct impact on the lexical chains. Their quality can be expected to be poor for resource-scarce languages or specialized application domains.

Statistical methods to modeling language semantics have proven to deliver good results in many natural language processing applications. In particular, probabilistic topic models have been successfully used for tasks such as summarization (Gong and Liu, 2001; Hennig, 2009), text segmentation (Misra et al., 2009), lexical substitution (Dinu and Lapata, 2010) or word sense disambiguation (Cai et al., 2007; Boyd-Graber et al., 2007).

Proceedings of NAACL-HLT 2013, pages 989–999,
Atlanta, Georgia, 9–14 June 2013.

In this work, we address the question, whether statistical methods for the extraction of lexical chains can yield better results than existing knowledge-based methods, especially for underresourced languages or domains, following principles of Structure Discovery (Biemann, 2012). To address this, we have developed a methodology for evaluating the quality of lexical chains intrinsically, have carried out an annotation study, and report results on a corpus of manually annotated German news documents.

After defining a measure for the comparison of (manually or automatically created) lexical chains in Section 2, Section 3 describes our annotation methodology and discusses issues regarding the inherent subjectivity of lexical chain annotation. In Section 4, three statistical approaches for lexical chaining are developed on the basis of the LDA topic model. Experiments that demonstrate the advantage of these approaches over a knowledge-baseline are conducted and evaluated in Section 5, and Section 6 concludes and provides an outlook future directions.

1.2 Previous Work on Lexical Chains

Morris and Hirst (1991) initially proposed an algorithm for lexical chaining based on Roget's thesaurus (Roget, 1852), and manually assessed the quality of their algorithm. Hirst and St-Onge (1998) first presented a computational approach to lexical chaining based on WordNet showing that the lexical database is a reasonable replacement to Roget's. The basic idea behind these algorithms is that semantically close words should be connected to form chains. Subsequent approaches mainly concentrated on disambiguation of words to WordNet concepts (WSD), since ambiguous words can lead to the overgeneration of connections. Barzilay and Elhadad (1997) improved the implicit word sense disambiguation (WSD) by keeping a list of different interpretations of the text and finally choosing the most plausible senses for chaining. Silber and McCoy (2002) introduced an efficient variant of the algorithm with linear complexity in the number of candidate terms. Galley and McKeown (2003) further improved accuracy by first performing WSD, and then using the remaining links between the disambiguated concepts only. They also introduced a so-called *disambiguation graph*, a representation that has also been utilized by the method of Medelyan (2007), where she applied a graph clustering algorithm to the disambiguation graph to cut weak links, performing implicit WSD. A combination of statistical and knowledge-based methods is presented by Marathe and Hirst (2010), who combine distributional co-occurrence information with semantic information from a lexicographic resource for extracting lexical chains and evaluate them by text segmentation. We are not aware of previous lexical chaining algorithms that do not rely on a lexicographic resource at all.

A major issue in developing a new lexical chaining algorithm is the comparison to previous systems. Most of previous approaches are validated by the evaluation in a certain task like summarization, word sense disambiguation, keyphrase extraction or information retrieval (Stairmand, 1996). Hence, these extrinsic evaluations are heavily influenced by the particular task at hand. We propose to re-consider lexical chaining as a task on its own, and propose objective criteria for directly comparing lexical chains to this end.

2 Comparing Lexical Chains

The comparison of lexical chains is a non-trivial task. We adopt the idea of interpreting lexical chains as clusters and a particular set of lexical chains as a clustering, and develop a suitable cluster comparison measure. As stated by Meilă (2005) and Amigó et al. (2009), a best clustering comparison measure for the general case does not exist. It should be stressed that the appropriate clustering measure highly depends on the task at hand.

After exploring a number of measures[1], we decided on a combination of the adjusted Rand index (ARI, Hubert and Arabie (1985)) and the *basic merge distance* (BMD, Menestrina et al. (2010)) for our new measure. Menestrina et al. (2010) introduced a linear time algorithm for computing the *generalized merge distance* (GMD), which counts

[1]Explored measures which are unsatisfactory for the given task are: Closest Cluster F_1 (Benjelloun et al., 2009), K (Ajmera et al., 2002), Pairwise F_1 (Manning et al., 2008), Variation of Information (Meilă, 2005), B^3(Bagga and Baldwin, 1998), V-Measure (Rosenberg and Hirschberg, 2007), Normalized Mutual Information (Strehl, 2002). The last two measures are equal. A proof of this can be found in the appendix.

split and *merge* cluster editing operations. Using a constant factor of 1 for both splits and merges gives the *basic merge distance* (BMD): Considering $\top$ as the most general clustering of a dataset D, where all elements are grouped into the same cluster, and further considering $\perp$ as the most specific clustering of D, where each element builds its own cluster, the *lattice* between $\top$ and $\perp$ spans all possible clusterings and the BMD can be interpreted as the *shortest path* from a clustering C to a clustering C' in the lattice with some restrictions (see Menestrina et al. (2010) for details). We normalize the BMD score by the maximum BMD[2] to the normalized basic merge distance ($NBMD$). ARI is is based on pair comparisons, and is computed as[3]:

$$index = TP$$

$$expected\ index = \frac{(TP+FP)\times(TP+FN)}{TP+TN+FP+FN}$$

$$max\ index = TP + \frac{1}{2}(FP+FN)$$

$$ARI(C,C') = \frac{index - expected\ index}{max\ index - expected\ index}$$

The reasons for choosing these two particular measures are the following: ARI is a well known measure which is adjusted (corrected) for decisions made by chance. But since it is based on pairwise element comparison it completely disregards singleton clusters (chains) and some types of errors are not adequately penalized. The NBMD on the other hand penalizes various errors almost equally.

We combine the two single measures into a new $lccm$ (lexical chain comparison measure), defined as the arithmetic mean between ARI and $1 - NBMD$. An $lccm$ of 1 indicates perfect congruence and an $lccm = 0$ indicates that not a single pair of items in C is found in a cluster together in C'.

$$lccm(C,C') = \frac{1}{2}\left[1 - NBMD(C,C') + ARI(C,C')\right].$$

[2] $BMD(\top,\perp)$ for $|D| \leq 2$, $BMD(\top,\perp)+1$ otherwise

[3] TP: pairs in D and D', FP: pairs in D' but not in D, FN: pairs in D but not in D', TN: pairs not in D and not in D', where D is the underlying dataset of C, D' is the underlying dataset of C', and pairs means all unique combinations of elements that are in the same cluster.

3 Annotating Lexical Chains

A challenge with the annotation of lexical chains is the subjective interpretation of the text by individual annotators (Morris and Hirst, 2004), which also substantiates the fact that currently no gold standard exist, and all previous automatic approaches are evaluated by performing a certain NLP task. Hollingsworth and Teufel (2005) as well as Cramer et al. (2008) conclude from their lexical chain annotation projects that high inter-annotator agreement is very hard to achieve. We argue that directly evaluating on lexical chains should enable us to optimize towards higher-quality chain annotations, which is a task of its own right and which has the potential to improve all subsequent applications. For this, we devise an annotation scheme that gets us reasonable inter-annotator agreement, inspired by the concept of *cohesive harmony* (Hasan, 1984), and report on an annotation project for German newswire texts.

Documents from the SALSA 2.0 (Burchardt et al., 2006) corpus were chosen to form the basis for the annotation of lexical chain information. SALSA is based on the semi-automatically annotated TIGER Treebank 2.1 (Brants et al., 2002). The TIGER treebank provides manual annotations, such as lemmas, part-of-speech tags, and syntactic structure, the SALSA part of the corpus is also partially annotated with FrameNet-style (Baker et al., 1998) frame annotation. The documents are general domain news articles from a German newspaper comprising about 1,550 documents and around 50,000 sentences in total, with a median document length of 275 tokens.

3.1 Annotation Scheme

In order to minimize the subjectiveness of choices by different annotators, annotation guidelines were developed comprising a total of ten pages. We decided to consider only nouns, noun compounds and non-compositional adjective noun phrases like "dirty money" as candidate terms for lexical chaining, which is consistent with the procedures of Hollingsworth and Teufel (2005) and Cramer et al. (2008). For annotation, we used the MMAX2[4] (Müller and Strube, 2006) tool.

We introduce the term *dense chain*, which refers to a type of lexical chain in which every element is

[4] `http://mmax2.sourceforge.net`

related to every other element in that chain. Terms are considered to be related if they share the same topic, i.e. common sense and knowledge of the language is needed to decide which terms belong together in the same topic and whether a chosen topic is neither too broad nor too narrow. A single dense chain can thus be assigned a definite topical description of its items. Whereas Hollingsworth and Teufel (2005) dealt with the inherent fuzziness of membership of terms to lexical chains by allowing terms to occur in different lexical chains, we follow the concept of *cohesive harmony* introduced by Hasan (1984) here, where complete chains can be linked to others. For this purpose, we introduce so-called *level two links*, which are cohesive ties between lexical items in distinct dense chains. Having such a link between two chains, both chains can be assigned a topical description which is broader than the description of the individual chains. This results in a two-level representation of chains. We report on dense lexical chains and merged lexical chains (dense chains are merged into a common chain if a level two link exists between them) separately.

In total, 100 documents were annotated by two expert annotators. Documents were chosen around the length median and consist of 248 – 304 tokens. The two rightmost columns of Table 3 show the characteristics of the annotated data set. It can be concluded that there is a moderate to high agreement regarding the annotator selections of candidate terms, which is ensured by preselection of candidate terms by part-of-speech patterns. A value of 81 % in the average agreement on lexical items (cf. Figure 1) shows that even though the choice of lexical items is limited to nouns and adjective noun phrases only, the decision on candidate termhood is somewhat different between the annotators, but compares favorably with previous findings of 63 % average pairwise agreement (Morris and Hirst, 2004).

Figure 2 shows the annotator agreement on the individual documents using the *lccm* (cf. Sec. 2), sorted in the same way as in Figure 1. In order to use the level two link information the figure also shows a second agreement score, which was computed on merged chains.

The agreement scores of the assignment of lexical items to lexical chains depend partially on the agreement scores of the identified lexical items themselves, which is a desired property. Across all documents, a perfect agreement was never achieved, which confirms the difficulty of annotating such a subjective task: The average $lccm$ per document on the manual annotations is 0.56 (dense chains), respectively 0.54 (merged chains). However, the considerable overlap between the annotators still enables us to evaluate automatic chaining methods, and the $lccm$ agreement score serves as an upper bound. Note that by performing no reconciliation of the annotations we explicitly allow the possibility of different interpretations which is in our opinion appropriate here due to the subjectiveness of the task itself. By doing so, we evaluate our algorithms against individual annotator interpretations.

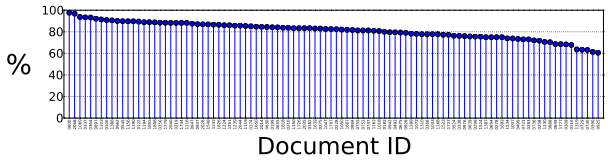

Figure 1: Agreement of lexical items annotated by annotator A and annotator B as a percentage of lexical items annotated by annotator A or annotator B. The average agreement is 81 %.

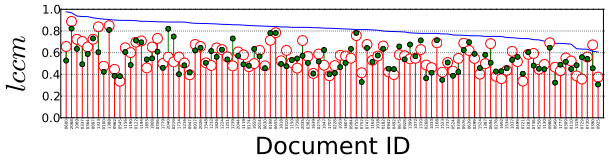

Figure 2: Individual annotator agreement scores on 100 documents sorted by their agreement on candidate terms. The red circles show the agreement of both annotators on the dense lexical chains disregarding the cohesive links, and the green dots show the agreement of both annotators on the merged lexical chains (via the cohesive links) both using the proposed *lexical chain comparison measure*.

4 Statistical Methods for Lexical Chaining

This work employs a well-studied statistical method for creating something that Barzilay (1997) called an *automatic thesaurus* which will then be adapted for lexical chaining. For our automatic approaches, candidate lexical items in a text are preselected by the same heuristic that is also applied in Section 3 for the annotation process.

Topic models (TMs) are a suite of unsuper-

vised algorithms designed for unveiling some hidden structure in large data collections. The key idea is that documents can be represented as composites of so-called *topics* where a topic itself represents as a composite of words. Hofmann (1999) defined a topic to be a probability distribution over words and a document to be a probability distribution over a fixed set of topics. We use the *latent Dirichlet allocation* (LDA, Blei et al. (2003)) topic model for estimating the semantic closeness of candidate terms, and explore different ways of utilizing LDA's topic information in automatic lexical chainers. Specifically, we use the GibbsLDA++[5] framework for topic model estimation and inference, and examine the following LDA parameters: number of topics T, Dirichlet hyperparameters for document-topic distribution α and topic-term distribution β.

We now describe three LDA-based approaches to lexical chaining.

4.1 LDA Mode Method (LDA-MM)

The LDA-MM approach places all word tokens that share the same topic ID into the same chain. The point is now how to decide to which topic a word belongs to. Since single samples of topics per word exhibit a large variance (Riedl and Biemann, 2012), we follow these authors by sampling several times and using the mode (most frequently assigned) topic ID per word as the topic assignment. This strategy reduced the variance in the $lccm$ to a tenth[6].

More formally, let $samples^{(d,w)}$ be the vector of assignments that have been collected for a certain word w in a certain document d with each $samples_i^{(d,w)}$ referring to the i-th sampled topic ID for (d, w). In other words, $samples^{(d,w)}$ can be seen as the Markov chain for a particular word in a particular document. Further let $z^{(d,w)}$ be the topic ID that was most assigned to the word w with respect to the samples in $samples^{(d,w)}$. Precisely, $z^{(d,w)}$ is defined to be the sampled mode in $samples^{(d,w)}$ — in case of multiple modes a random mode is chosen, which never happened in our experiments.

$$\begin{aligned} z^{(d,w)} &= \text{mode}\left(samples^{(d,w)}\right) \\ &\approx \arg\max_j \left(P(z = j|w, d)\right) \end{aligned}$$

The LDA-MM assigns for every word w which is a candidate lexical item of a certain document d which is assigned the same topic $z^{(d,w)}$ to the same chain; hence implicitly disambiguating the terms.

The possibility to create level two links is given by taking the second most occurring topic for a given word if it exceeds a certain threshold.

4.2 LDA Graph Method (LDA-GM)

The LDA-GM algorithm creates a similarity graph based on the comparison of topic distributions for given words and then applies a clustering algorithm in order to find semantically related words.

Let $\psi^{(d,w)}$ be the per-word topic distribution $P(z|w, d)$. Analogously to the LDA-MM, $\psi^{(d,w)}$ can be obtained by counting the occurrences of a certain topic ID z in the sample collection $samples^{(d,w)}$ for a particular word w and document d.

The semantic relatedness between any two words w_i and w_j can then be measured by their similarity score of the topic distributions $\psi^{(d,w_i)}$ and $\psi^{(d,w_j)}$, which is stored in a term similarity matrix. This matrix can also be interpreted as an adjacency matrix of a graph, with candidate items being nodes and edges being weighted with the similarity value sim_{ij} for any two nodes $i, j : i \neq j \ \wedge \ i, j \in \{1, 2, \ldots, N_d\}$. We test two similarity measures: Euclidian (dis-)similarity and cosine similarity.

Let $G = (V, E)$ be the graph representation of a document with term vertices $V = \{v_1, \ldots, v_{Nd}\}$ and weighted edges $E = \{(v_1, v_2, sim_{12}), \ldots (v_{Nd}, v_{Nd-1}, sim_{N_d N_d-1})\}$, where sim_{ij} is either the cosine or Euclidean similarity of term vectors. For simplicity, we reduce this representation to an unweighted graph by only retaining edges (of unit weight) that have a similarity above a parameter threshold ϵ_{sim}. To identify chains as clusters in this graph, we follow Medelyan (2007) and apply the *Chinese Whispers* graph clustering algorithm (CW, Biemann (2006)), which finds the number of clusters automatically. The CW algorithm implementation comes with

[5] `http://gibbslda.sourceforge.net`

[6] Preliminary experiments yielded a variance of 2.6×10^{-6} in $lccm$ using the mode method and 3.07×10^{-5} using a single sample for lexical chain assignment.

three parameters to regulate the node weight based on its degree, which influences cluster size and granularity. We test options "top", "dist log" and "dist lin".

The final chaining procedure is straightforward: The LDA-GM algorithm assigns every candidate lexical item w_i of a certain document d which is assigned the same class label c_i to the same chain. Level two links are drawn using the second dominant class of a vertex's neighborhood, which is provided by the CW implementation.

4.3 LDA Top-N Method (LDA-TM)

The LDA-TM method is different to the others in that it uses the information of the per-topic word distribution $\phi^{(z)} = P(w|z)$ and the per-document topic distribution $\theta^{(d)} = P(z|d)$. Given a parameter n referring to the top n topics to choose from $\theta^{(d)}$ and a parameter m referring to the top m words to choose from $\phi^{(z)}$ the main procedure can be described as follows: for all $z \in$ top n topics in $\theta^{(d)}$: chain the top m words in $\phi^{(z)}$.

Note that although the number of chains and chain members for each chain is bound and could lead to the same number and sizes of chains, in practice the number of generated chains as well as the number of chain members still varies considerably across documents: often some of the top m words for a (globally computed) topic do not even occur in a particular document. This implies that the parameters n and m must not be set globally but dependent on the particular document. To overcome this to some extent, additional thresholding parameters ϵ_θ and ϵ_ϕ are used for further bounding the respective n or m parameter. The procedure works like this: for all z $\in$ top n topics in $\theta^{(d)} \ \wedge \ \theta_z^{(d)} < \epsilon_\theta$: chain the top m words w in $\phi^{(z)} \ \wedge \ \phi_w^{(z)} < \epsilon_\phi$.

Level two links are created by computing the cosine similarity between every pair of the top n topic distributions, and thresholding with a link parameter.

4.4 Repetition Heuristic

All methods described above can be applied to new unseen documents that are not in the training set. To alleviate a possible vocabulary mismatch between training set and test set, which happens when terms in the test set have not been contained in our training documents, we add a heuristic that chains repetitions of (previously unknown) words as a post-processing step to all methods.

5 Empirical Analysis

In order to provide a realistic estimate of the quality of our methods to unseen material, we randomly split our annotated documents in two parts of 50 documents each. One part is used as a *development set* for optimizing the parameters of the methods (i.e. model selection), the other part forms our *test set* for evaluation.

The *training corpus*, on the other hand, consists of all 1,211 SALSA/Tiger documents that are not part of the development and test corpus and neither very long nor very short. These documents are taken from the German newspaper "Frankfurter Rundschau" around 1992. Additionally the training corpus is enriched with 12,264 news texts from the same newspaper around 1997 with similar characteristics[7], making up a total of 13,457 training documents for the estimation of topic models.

Input to the LDA model training are verbs, nouns and adjectives, as well as candidate terms as described in Section 3.1, all in their lemmatized form. We further filter words that occur in more than 1/3 of the training documents, as well as known stopwords, and words that occur in less than two documents which results in a vocabulary size of about 100K words.

5.1 Experimental Setup

For comparison, we implemented three baselines, which we describe below. One baseline is trivial, two baselines are state-of-the art knowledge-based systems adapted to German.

Random: Candidate lexical items are randomly tied together to form sets of lexical chains. Level two links are created analogously. We regulate the process to yield the same average number of chains and links as in the development and test data.

S&M GermaNet: Algorithm by Silber and McCoy (2002) with GermaNet as its knowledge resource.

[7] as provided by Projekt Deutscher Wortschatz, http://wortschatz.uni-leipzig.de/

G&M GermaNet: Algorithm by Galley and McKeown (2003), also using GermaNet.

GermaNet (Hamp and Feldweg, 1997) is a large WordNet-like resource for German, containing almost 100,000 lexical units and over 87,000 conceptual relations between synsets. While its size is only about half of WordNet, it is one of the largest non-English lexical semantic resources.

5.2 Model Selection

We optimize two sets of parameters: parameters for the LDA topic model (number of topics K, Dirichlet hyperparameters α and β) are optimized for the LDA-MM method only, and the same LDA model is used in the other two LDA-based methods. Parameters particular to the respective method are optimized individually. For LDA, we tested sensible combinations in the ranges $K = 50..1000$, $\alpha = 0.05/K..50/K$ and $\beta = 0.001..0.1$. The highest performance of the LDA-MM method was found for $K = 500$, $\alpha = 50/K$, $\beta = 0.001$, and the resulting topic model is used across all methods. The final parameter values for the other methods, found by exhaustive search, are summarized in Table 1.

Method	Parameter
LDA-GM	$similarityfunction$ = cosine similarity $labelweightscheme$ = dist log ϵ_{sim} = 0.95
LDA-TM	n = 10, m = 20, ϵ_θ = 0.2, ϵ_ϕ = 0.2

Table 1: Final parameter values.

5.3 Evaluation

For evaluation purposes, terms that consist of multiple words are mapped to its rightmost term which is assumed to be the head, e.g. "dirty money" is mapped to "money". Additionally, singleton chains, i.e. chains that contain only a single lexical item are omitted unless the respective lexical item is not linked by a level two link.

Dense Chains Comparative results of the approaches in terms of $lccm$ for both annotators are summarized in Table 2 (upper half). We observe that all our new methods beat the random baseline and the two knowledge-based baselines by a large margin. The knowledge-based baselines, both using GermaNet, produce very similar $lccm$ scores, which highlights the important role of the knowledge resource. Data analysis revealed that while chains produced by knowledge-based baselines are sensible, the main problem is a lack of coverage in terms of vocabulary and relations in GermaNet. Comparing the statistical methods, the LDA-GM method excels over the others.

	Anno A	Anno B	Average
LDA-MM	0.320	0.306	0.313
LDA-TM	0.307	0.299	0.303
LDA-GM	**0.328**	**0.314**	**0.321**
G&M	0.255	0.215	0.235
S&M	0.248	0.209	0.229
Random	0.126	0.145	0.135
LDA-MM	**0.316**	**0.300**	**0.308**
LDA-TM	0.303	0.280	0.291
LDA-GM	0.279	0.267	0.273
G&M	0.184	0.166	0.176
S&M	0.179	0.159	0.169
Random	0.196	0.205	0.201

Table 2: Results of the evaluation based on dense chains (upper half) and merged chains (lower half). The annotator agreement on the test set's chains = 0.585; on merged chains = 0.553

Level Two Links Table 2 (lower half) summarizes the evaluation results of the merged chains via level two links. Because of merging, a text now contains fewer chains with more lexical items each. Note that knowledge-based baselines do not construct level two links, which is why they are heavily penalized in this setup.

Again, the statistical methods beat the baselines by a substantial amount. In this evaluation, the random baseline performs above the knowledge-based methods, which is rooted in the fact that $lccm$ penalizes small, correct chains, whereas the random baseline with linking often produces very large chains containing most of the terms – something that we also observe for many manually annotated documents. The large overlap in the biggest chain then leads to the comparatively high random baseline score. In this evaluation, the LDA-MM is the clear winner, with LDA-GM being clearly inferior this time.

	LDA-MM	LDA-GM	LDA-TM	S&M	G&M	Anno A	Anno B
avg. num. of lexical items per doc.	38.20	29.32	30.82	14.40	15.29	38.66	38.96
avg. num. of chains per doc.	13.80	9.12	7.32	5.83	5.71	11.25	7.38
avg. num. of links per doc.	8.60	2.06	1.44	–	–	5.47	2.41
avg. size lexical chains	2.82	3.41	4.61	2.48	2.68	3.69	5.57
avg. num. of merged lexical chains	5.76	7.06	5.98	–	–	6.10	4.99
avg. size merged lexical chains	8.29	4.45	5.57	–	–	7.60	8.91

Table 3: Quantitiative characteristics of automatic and manual lexical chains. In average, a document contains 51.58 candidate terms as extracted by our noun phrase patterns

Davud Bouchehri, seit der letzten ***Spielzeit*** *als* ***Dramaturg*** *in Basel tätig, wechselt zur*
[Davud Bouchehri,] [since] [the] [last] [playing period] [as] [dramaturg] [in] [Basle] [acting,] [switches] [to the]
Saison *1996 / 97 als* ***künstlerischer Geschäftsführer*** *des* ***Schauspiels*** *an das* ***Staatstheater*** *Darmstadt. Der*
[1996 / 97 season] [as] [art] [director] [of the] [play] [to] [the] [state theater] [Darmstadt.] [The]
aus dem Iran stammende 34jährige soll daneben auch für ***spartenübergreifende Projekte*** *zuständig sein,*
[from] [the] [Iran] [coming] [34-year-old] [shall] [besides] [also] [for] [multi discipline] [projects] [responsible] [be,]
teilte das Basler ***Theater*** *am Donnerstag mit.*
[aquainted] [the] [Basle's] [theather] [on] [Thursday] [with.]

LDA-MM:
c_1: {*Spielzeit, Schauspiels, Staatstheater*}
c_2: {*Dramaturg, Theater*}
c_3: {*Saison*}
l_1: *(Theater → Spielzeit)*
l_2: *(Spielzeit → Saison)*

LDA-GM:
c_1: {*Dramaturg, Theater*}
c_2: {*Schauspiels, Staatstheater*}

LDA-TM:
c_1: {*Schauspiels, Staatstheater, Theater*}
c_2: {*Dramaturg*}
c_3: {*Spielzeit, Saison*}
l_1: *(Theater → Dramaturg)*

S&M-GermaNet:
–

G&M-GermaNet:
c_1: {*Staatstheater, Theater*}

Figure 3: Diverse output of the various lexical chaining systems after applying them on a short German example text from the used TIGER/SALSA corpus. For a better understanding the text is calqued. Candidate items are highlighted and the c_i are the resulting dense lexical chains and the l_i are the level two links produced by the various methods.

Data Analysis Table 3 shows quantitative numbers of the extracted lexical chains in the test set.

The LDA-MM approach chains and links a lot more items than the other statistical methods: it creates a lot more links between items that would otherwise be removed because they form unlinked singleton chains. As opposed to this, the graph method (LDA-GM), as well as the top-n method (LDA-TM) perform an implicit filtering on the candidate lexical items by creating less level two links, yet larger dense chains. The knowledge based algorithms by Silber and McCoy (2002) and Galley and McKeown (2003) extract fewer and smaller chains than the statistical approaches, which reflects GermaNet's sparsity issues. While higher lexical coverage in the underlying resource would increase the coverage of our knowledge-based systems, this is only one part of the story. The other part is rooted in the fact that lexical cohesion relations, which are used in lexical chains, encompass many more semantic relations than listed in today's lexical semantic networks. This especially holds for cases where several expressions refer to the same event or theme for which no well-defined relation exists, such as e.g. "captain" and "harbor".

Comparing the three LDA-based approaches, no overall best method could be determined. the LDA-MM seems especially suited for a high coverage and coarse (level two) chains, the LDA-GM appears most suited for dense chains, and LDA-TM produces the longest chains on average.

Figure 3 shows the resulting dense lexical chains and level two links after applying our chainers to a short example text from our corpus. In the example the LDA-TM produces the most adequate lexical chains, at least in our intuition. The LDA-GM and the LDA-MM produce slightly wrong chains, yet the LDA-MM additionally creates some meaningfull level two links which the LDA-GM does not. Both knowledge-based approaches perform poorly compared to the knowledge-free approaches, where the S&M algorithm creates no chains at all and the

G&M algorithm produces only a single chain containing only two words. This is mostly due to GermaNet's lacking lexical and relational coverage and the scope of the algorithms for finding relations between the words.

6 Conclusion

In this paper, we presented experiments for automatic lexical chain annotation and evaluated them directly on a manually annotated dataset for German. A new two-level annotation scheme for lexical chains was proposed and motivated by the concept of cohesive harmony. We further proposed a new measure for comparing lexical chain annotations that is especially suited for the characteristics of lexical chain annotations. Three variants of statistical lexical chaining methods based on the LDA topic model were proposed and evaluated against two knowledge-based baseline systems. Our statistical methods exhibit a substantially higher performance than the knowledge-based systems on our dataset. This can partially be attributed to missing relations, partially to the lack of lexical coverage of GermaNet, which was used in these systems. Since GermaNet is a large lexical-semantic net, however, this strengthens our main point: Especially for under-resourced languages or subject domains, statistical and data-driven methods should be preferred over their knowledge-based counterparts, since they do not require the development of lexical-semantic nets and adopt easily to subject domains by training their unsupervised models on an in-domain collection.

In future work, we would like to explore better ways of selecting candidate items. While our POS-pattern-based selection mechanism works for practical purposes, it currently only extracts noun phrases and over-generates on compositional adjective modifiers. We would like to define a better filter to reduce over-generation. Further, especially for compounding languages such as German, we would like to decompose one-word compounds as to be able to link their heads in lexical chains.

While we found it important to directly evaluate our lexical chaining algorithms on manually annotated data, a natural next step in this line of research is to use our lexical chaining methods as pre-processing steps for applications such as summarization, text segmentation or word sense disambiguation. This would enable to find out advantages and disadvantages of our three variants with respect to an application.

The manually annotated data, the open source annotation tool, the annotation guidelines and the implementations of all described methods and baselines are available for download[8].

Acknowledgments

This work has been supported by the Hessian research excellence program *Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE)* as part of the research center *Digital Humanities*.

Proof: Equality of NMI and V

Using the standard notation from information retrieval $H(X)$= Entropy, $I(X,Y)$= Information, $H(X|Y)$= Conditional Entropy, NMI(X,Y)= Normalized Mutual Information, V(X,Y)= V-Measure:

$$V(C,K) = 2 \times \frac{h \times c}{h + c} \tag{1}$$

$$h = 1 - \frac{H(C|K)}{H(C)} \quad , \quad c = 1 - \frac{H(K|C)}{H(K)} \tag{2}$$

and

$$NMI(C,K) = \frac{I(C,K)}{\frac{H(C)+H(K)}{2}} = 2 \times \frac{I(C,K)}{H(C)+H(K)} \tag{3}$$

reformulate h and c using the fact that $I(C,K) = H(C) - H(C|K) = H(K) - H(K|C)$:

$$\begin{aligned} h &= 1 - \frac{H(C|K)}{H(C)} \\ &= \frac{H(C)}{H(C)} - \frac{H(C|K)}{H(C)} \\ &= \frac{I(C,K)}{H(C)} \end{aligned} \tag{4}$$

$$\begin{aligned} c &= 1 - \frac{H(K|C)}{H(K)} \\ &= \frac{H(K)}{H(K)} - \frac{H(K|C)}{H(K)} \\ &= \frac{I(C,K)}{H(K)} \end{aligned} \tag{5}$$

simplifying $h \times c$ using (4) and (5):

$$\begin{aligned} h \times c &= \frac{I(C,K)}{H(C)} \times \frac{I(C,K)}{H(K)} \\ &= \frac{I(C,K)^2}{H(C)H(K)} \end{aligned} \tag{6}$$

simplifying $h + c$ using (4) and (5):

$$\begin{aligned} h + c &= \frac{I(C,K)}{H(C)} + \frac{I(C,K)}{H(K)} \\ &= \frac{I(C,K)H(K) + I(C,K)H(C)}{H(C)H(K)} \\ &= \frac{I(C,K)[(H(K) + H(C)]}{H(C)H(K)} \end{aligned} \tag{7}$$

simplifying $\frac{h \times c}{h+c}$ using (6) and (7):

$$\begin{aligned} \frac{h \times c}{h + c} &= \frac{I(C,K)^2}{H(C)H(K)} \times \frac{H(C)H(K)}{I(C,K)[H(K) + H(C)]} \\ &= \frac{I(C,K)}{H(K) + H(C)} \end{aligned} \tag{8}$$

[8] http://www.ukp.tu-darmstadt.de/data/lexical-chains-for-german/

substituting (8) into (1) shows that NMI and V are equal:

$$V(C,K) = 2\times\frac{h\times c}{h+c} = 2\times\frac{I(C,K)}{H(K)+H(C)} = NMI(C,K) \quad (9)$$

References

Jitendra Ajmera, Hervé Bourlard, and I. Lapidot. 2002. Unknown-Multiple Speaker clustering using HMM. In *Proceedings of the International Conference of Spoken Language Processing*, ICSLP '02, Denver, Colorado, USA.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12:461–486.

Amit Bagga and Breck Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation*, LREC '98, pages 563–566, Granada, Spain.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *COLING '98: Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90, Montreal, Quebec, Canada.

Regina Barzilay and Michael Elhadad. 1997. Using Lexical Chains for Text Summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain.

Regina Barzilay. 1997. Lexical Chains for Summarization. Master's thesis, Ben-Gurion University of the Negev, Beersheva, Israel.

Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Whang, and Jennifer Widom. 2009. Swoosh: a generic approach to entity resolution. *The VLDB Journal*, 18:255–276.

Chris Biemann. 2006. Chinese Whispers – an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing. In *Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80, New York City, USA.

Chris Biemann. 2012. *Structure Discovery in Natural Language*. Theory and Applications of Natural Language Processing. Springer Berlin / Heidelberg.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A Topic Model for Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033, Prague, Czech Republic.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT02)*, Sozopol, Bulgaria.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA Corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th international conference on Language Resources and evaluation (LREC-2006)*, Genoa, Italy.

Junfu Cai, Wee Sun Lee, and Yee Whye Teh. 2007. Improving Word Sense Disambiguation Using Topic Features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1015–1023, Prague, Czech Republic.

Joe Carthy. 2004. Lexical Chains versus Keywords for Topic Tracking. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2945 of *Lecture Notes in Computer Science*, pages 507–510. Springer, Berlin / Heidelberg.

Irene Cramer, Marc Finthammer, Alexander Kurek, Lukas Sowa, Melina Wachtling, and Tobias Claas. 2008. Experiments on Lexical Chaining for German Corpora: Annotation, Extraction, and Application. *Journal for Language Technology and Computational Linguistics (JLCL)*, 23(2):34–48.

Georgiana Dinu and Mirella Lapata. 2010. Topic Models for Meaning Similarity in Context. In *COLING '10: Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 250–258, Beijing, China.

Michel Galley and Kathleen McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *IJCAI'03: Proceedings of the 18th international joint conference on Artificial intelligence*, pages 1486–1488, Acapulco, Mexico.

Yihong Gong and Xin Liu. 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–25, New Orleans, Louisiana, USA.

Stephen J. Green. 1996. Using Lexical Chains to Build Hypertext Links in Newspaper Articles. In *AAAI-96 Workshop on Internet-based Information Systems*, pages 115–141, Portland, Oregon, USA.

Michael A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. English language series. Longman, London.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceed-*

ings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, Spain.

Ruqaiya Hasan. 1984. Coherence and Cohesive Harmony. In James Flood, editor, *Understanding Reading Comprehension*, Cognition, Language, and the Structure of Prose, pages 181–220. International Reading Association, Newark, Delaware, USA.

Leonhard Hennig. 2009. Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Proceedings of the International Conference RANLP-2009*, pages 144–149, Borovets, Bulgaria.

Graeme Hirst and David St-Onge. 1998. Lexical Chains as representation of context for the detection and correction malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, Language, Speech, and Communication, pages 305–332. The MIT Press, Cambridge, Massachusetts, USA.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI '99, pages 289–296, Stockholm, Sweden.

William Hollingsworth and Simone Teufel. 2005. Human annotation of lexical chains: Coverage and agreement measures. In *Proceedings of the Workshop ELECTRA: Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications*, In Association with SIGIR '05, Salvador, Brazil.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Meghana Marathe and Graeme Hirst. 2010. Lexical Chains Using Distributional Measures of Concept D. In *Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'10, pages 291–302, Iaşi, Romania.

Olena Medelyan. 2007. Computing lexical chains with graph clustering. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, ACL '07, pages 85–90, Prague, Czech Republic.

Marina Meilă. 2005. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pages 577–584, Bonn, Germany.

David Menestrina, Steven Euijong Whang, and Hector Garcia-Molina. 2010. Evaluating entity resolution results. *Proceedings of the VLDB Endowment*, 3(1):208–219.

Hemant Misra, François Yvon, Joemon Jose, and Olivier Cappé. 2009. Text Segmentation via Topic Modeling: An Analytical Study. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM 2009, pages 1553–1556, Hong Kong, China.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48.

Jane Morris and Graeme Hirst. 2004. The Subjectivity of Lexical Cohesion in Text. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Palo Alto, California, USA.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Manabu Okumura and Takeo Honda. 1994. Word sense disambiguation and text segmentation based on lexical cohesion. In *COLING '94: Proceedings of the 15th Conference on Computational Linguistics*, volume 2, pages 755–761, Kyoto, Japan.

Martin Riedl and Chris Biemann. 2012. Sweeping through the Topic Space: Bad luck? Roll again! In *ROBUS-UNSUP 2012: Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP held in conjunction with EACL 2012*, pages 19–27, Avignon, France.

Peter Mark Roget. 1852. *Roget's Thesaurus of English Words and Phrases*. Longman Group Ltd., Harlow, UK.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic.

H. Gregory Silber and Kathleen F. McCoy. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4):487–496.

Mark A. Stairmand. 1996. *A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval.* Ph.D. thesis, Center for Computational Linguistics, UMIST, Manchester.

Nicola Stokes, Joe Carthy, and Alan F. Smeaton. 2004. SeLeCT: A Lexical Cohesion Based News Story Segmentation System. *AI Communications*, 17(1):3–12.

Alexander Strehl. 2002. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. Ph.D. thesis, University of Texas, Austin.