Comparison between LSA-LDA-Lexical Chains

Costin Chiru, Traian Rebedea and Silvia Ciotec
University Politehnica of Bucharest, Department of Computer Science and Engineering,
313 Splaiul Independetei, Bucharest, Romania
{costin.chiru, traian.rebedea}@cs.pub.ro, silvia.ciotec@gmail.com

Keywords: Latent Semantic Analysis - LSA, Latent Dirichlet Allocation - LDA, Lexical Chains, Semantic Relatedness.

Abstract: This paper presents an analysis of three techniques used for similar tasks, especially related to semantics, in Natural Language Processing (NLP): Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and lexical chains. These techniques were evaluated and compared on two different corpora in order to highlight the similarities and differences between them from a semantic analysis viewpoint. The first corpus consisted of four Wikipedia articles on different topics, while the second one consisted of 35 online chat conversations between 4-12 participants debating four imposed topics (forum, chat, blog and wikis). The study focuses on finding similarities and differences between the outcomes of the three methods from a semantic analysis point of view, by computing quantitative factors such as correlations, degree of coverage of the resulting topics, etc. Using corpora from different types of discourse and quantitative factors that are task-independent allows us to prove that although LSA and LDA provide similar results, the results of lexical chaining are not very correlated with neither the ones of LSA or LDA, therefore lexical chains might be used complementary to LSA or LDA when performing semantic analysis for various NLP applications.

1 INTRODUCTION

Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997), Latent Dirichlet Allocation (LDA) (Blei et. al, 2003) and lexical chains (Halliday and Hasan, 1976; Morris and Hirst, 1991) are widely used in NLP applications for similar tasks. All these methods use semantic distances or similarities/ relatedness between terms to form topics or chains of words. LSA and LDA use the joint frequency of the co-occurrence of words in different corpora, while the lexical chains technique uses WordNet (http://wordnet.princeton.edu/) synsets and links between them to find groups of highly-connected or closely-related words.

Although these methods can be similarly used for various NLP tasks - text summarization (Barzilay and Elhadad, 1997; Gong and Liu, 2001; Haghighi and Vanderwende, 2009), question answering (Novischi and Moldovan, 2006) or topic detection (Carthy, 2004) - they calculate different measures, having different meanings. LDA generates topical threads under a prior Dirichlet distribution, LSA produces a correlation matrix between words and documents, while lexical chains use the WordNet structure to establish a connection between synsets.

Therefore, the comparison and interpretation of similarities and differences between the aforementioned methods is important to understand which model might be the most appropriate for a given scenario (task and discourse type, for example). Previous studies were aimed at comparing different similarity measures built on top of WordNet in order to decide which one gives better results (Barzilay and Elhadad, 1997), or to compare the results provided by the lexical chains built using different measures with the ones given by LSA in order to add a further relationship layer to WordNet for improving its usefulness to NLP tasks (Boyd-Graber et. al, 2006). However, more recently Cramer (2008) pointed out that the existing studies are inconsistent to each other and that human judgments should not be used as a baseline for the evaluation or comparison of different semantic measures.

This work aims to study the behaviour of the three methods: LSA, LDA and lexical chains, based on a series of tests performed on two corpora: one consisting on four Wikipedia articles on different topics and another one built from multi-party online chat conversations debating four pre-imposed topics: forum, chat, blog, wikis.

The paper continues with a review of the evaluated techniques. Afterwards, we present the procedure for comparing the three methods along with the texts used for evaluation. Section 4 describes the obtained results and our observations, while the last section highlights the main conclusions of the study.

2 EVALUATED METHODS

2.1 LSA – Latent Semantic Analysis

LSA (Landauer and Dumais, 1997) is a statistical method for extracting the relations between words in texts. It is a corpus-based method that does not use dictionaries, semantic networks, grammars, syntactic or morphological parsers, and its input is represented only by raw text divided in "*chunks*". A chunk may be a sentence, an utterance in a chat, a paragraph or even a whole document, depending on the corpus. The method starts from the term-doc matrix computed on the corpus segmented into chunks and then applies a singular value decomposition in order to compute the most important singular values. Then, it produces a representation in a new space, called the latent semantic space, which uses only the most important (large) k singular values. The value for k depends on the corpus and task, and is usually between 100 and 600, a common choice being 300. This new space is used to compute similarities between different words and even whole documents, practically considering that words that are co-occurring in similar contexts may be considered to be semantically related.

2.2 LDA – Latent Dirichlet Allocation

LDA (Blei et. al, 2003) is a generative probabilistic model designed to extract topics from text. The basic idea behind LDA is that documents are represented as random mixtures of latent topics, where each topic is characterized by a set of pairs word-probability, representing the probability that a word belongs to a topic.

LDA assumes the following generative process for each document in a corpus: for each word $w_{d,i}$ in the corpus, it generates a topic z dependent on the mixture θ associated to the document d and then it generates a word from the topic z. To simplify this basic model, the size of the Dirichlet distribution k (the number of topics z) is assumed to be known and fixed. The Dirichlet prior is used because it has several convenient properties that facilitate inference and parameter estimation algorithms for LDA.

2.3 Lexical Chains

Lexical chains are groups of words that are semantically similar (Halliday and Hasan, 1976; Morris and Hirst, 1991). Each word in the chain is linked to its predecessors through a certain lexical cohesion relationship. Lexical chains require a lexical database or an ontology (most of the time, this database is WordNet) for establishing a semantic similarity between words. For this task, we have used WordNet and the Jiang-Conrath measure (Jiang and Conrath, 1997). As this measure requires the frequency of words in the English language and since we didn't have access to a relevant corpus, we have used the number of hits returned by a Google search for each of the considered words. Once the distances between words were computed, we have used a full-clustering algorithm to group the words in chains. The algorithm worked in an online fashion (each word was evaluated in the order of their appearance in the analyzed text), adding a word to an existing cluster only if it was related to more than 90% of the words that were already part of that chain. If the considered word could not be fitted in any of the existing chains, then we created a new chain containing only that specific word (Chiru, Janca and Rebedea, 2010).

3 COMPARISON METHODOLOGY

Experiments were conducted on two different corpora:

- a corpus composed of four articles from Wikipedia that were debating completely different topics: *graffiti*, *tennis*, *volcano* and *astrology,* consisting of 294 paragraphs and having a vocabulary size of 7744 words. In order not to have our results affected by noise, we removed from the corpus pronouns, articles, prepositions and conjunctions.
- a corpus consisting of 35 online chat conversations debating four pre-imposed topics: forum, chat, blog, wikis, each of them involving between 4 to 12 participants. This corpus consisted of 6000 utterances (41902 words), with a vocabulary size of 2241 words.

3.1 Methods for Obtaining the Results

The SVD is performed using the airhead-research package (https://code.google.com/p/airhead-research/wiki/LatentSemanticAnalysis) and a value of k = 300. Then, the LSA results are obtained starting from the matrix of similarities between each pair of words in the corpus. The degree of similarity between two words is computed using the cosine of the corresponding vectors in the latent space.

For LDA, the results are obtained from the distribution of each topic's words and the corresponding probabilities. In the first corpus, containing encyclopaedic articles from four different domains, we decided to use a number of topics $k = 4$ for this analysis. For the second corpus, consisting on debates on four imposed topics, we decided to use $k = 5$ topics for the analysis, as besides the imposed topics, the participants also inputted some off-topic content that could have been considered as the fifth topic. In order to better understand the behaviour of LDA, we extracted the top 35, 50, 100, 150 and 200 words that were considered representative for each topic, given that each article contained over 1000 words. The topic models were extracted using MALLET - MAchine Learning for LanguagE Toolkit (http://mallet.cs.umass.edu/).

In the case of lexical chains, we analyzed the words from each chain and also considered the maximum length and the total number of the lexical chains from a document (chat or Wikipedia article).

3.1.1 LDA - LSA Comparison

In order to compare the two methods, we started from the LDA topics and computed an LSA score for each concept from each topic generated by LDA. This score represented the average similarity between the target concept and each of the remaining words from the topic. The assessment of the relationship between LSA and LDA scores distributions was performed using Pearson's correlation coefficient and Spearman's rank correlation coefficient. LSA and LDA have also been compared on several NLP tasks, such as predicting word associations (Griffiths et al., 2007) and automatic essay grading (Kakkonen et al., 2008).

3.1.2 LSA - Lexical Chains Comparison

For comparing these two methods, we determined a similarity value for each lexical chain based on the LSA similarity as follows: we computed the LSA similarity between any pair of two words from the chain and averaged over all the words in that chain. LSA has been previously compared with semantic distances in WordNet (Tsatsaronis et al., 2010), but not with lexical chains.

3.1.3 LDA - Lexical Chains Comparison

This comparison is based on the number of common words between the lexical chains and the LDA topics. For each LDA topic we extracted a number of 35, 50, 100, 150 and 200 words, and computed different statistics for each case. To our knowledge, LDA and lexical chains have only been compared as an alternative for text segmentation (Misra et al., 2009).

4 EXPERIMENTAL RESULTS

4.1 Wikipedia Corpus

4.1.1 LDA - LSA Comparison

Table 1 presents the top 10 words from the 4 LDA topics of the first corpus. In Table 2 we present the most similar 30 word-pairs generated by LSA. We need to mention that LSA was trained on the concatenation of all 4 articles from Wikipedia.

Table 1: Top 10 words from the LDA topics for the Wikipedia corpus.

Topic 0	Topic 1	Topic 2	Topic 3
graffiti	tennis	volcanoes	astrology
new	game	volcano	been
culture	player	lava	Chinese
form	first	volcanic	personality
york	players	surface	scientific
design	two	example	based
popular	court	formed	considered
hip	three	examples	birth
style	point	extinct	bce
spray	French	flows	belief

Table 2: Top 30 most similar word-pairs generated by LSA for the Wikipedia corpus.

LSA Word Pairs		
men-cup	mid-thinning	plates-tectonic
mid-crust	tie-addition	center-baseline
hop-music	thinning-ridge	choice-receiver
mid-ridge	pace-receiver	depicted-dealer
lake-park	shift-equinox	degrees-equinox
mm-bounce	basque-perera	gladiatorial-cil
lady-week	degrees-shift	difficult-extinct
são-brazil	rhode-newport	tectonic-ridge
force-hero	federation-itf	era-compete
test-results	mud-formation	lifespans-volcanologist

For each topic we plotted the distributions of LDA and LSA scores for each word from that topic, computed as described in the previous section. Each LDA topic has 35 words that are sorted decreasing according to the LSA scores. The best result we have obtained was for the Topic 1 (*tennis*), where with very few exceptions, the LSA and LDA scores were very well correlated (0.855). This case is presented in Figure 1, where the x-axis represents the word number from the LDA topic and on the y axis we plotted the LDA and LSA scores corresponding to that word. The words' probabilities for the considered topic computed with LDA are represented by the blue colour while in red we present the LSA scores. The scattering diagram for the same topic is presented in Figure 2.

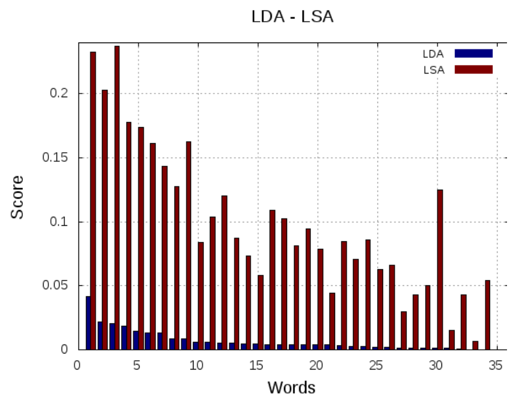

Figure 1: LDA – LSA distributions for Topic 1 (tennis) from the Wikipedia corpus.

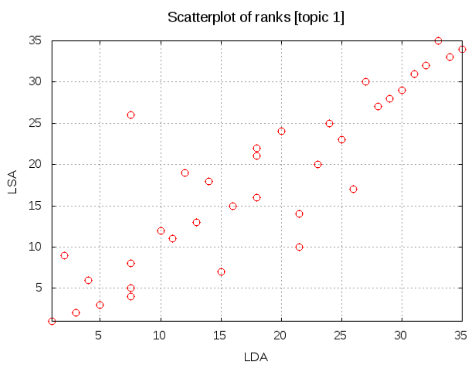

Figure 2: Scattering plot for the rank distributions for the LDA – LSA comparison for Topic 1 (tennis).

For a better visualization of the relationship between the two distributions, we present in Table 3 the Pearson's correlation and the Spearman's rank correlation coefficients between the LDA and LSA scores for each of the four LDA topics. With one exception, these values are close to 1, indicating a very good correlation (the strongest is highlighted in bold).

Table 3: LDA-LSA Pearson's Coefficient for the Wikipedia corpus.

Topic	Pearson's Coefficient	Spearman's Coefficient
0 (graffiti)	0.560	0.778
1 (tennis)	**0.855**	**0.873**
2 (volcanoes)	0.782	0.840
3 (astrology)	0.745	0.745

These results prove that there is clearly a correlation between the two distributions because both tend to decrease towards the last words of the topic. However, there are some words for which the two scores are discordant. We have extracted them and obtained the following results:

- for Topic 0 (graffiti): hip, produced, styles, non, offered, property;
- for Topic 1 (tennis): point, receiving;
- for Topic 2 (volcanoes): extinct, gases, features, falls;
- for Topic 3 (astrology): considered, challenge, avoid.

It is interesting to observe that the better correlated the LSA and LDA scores are for a given topic, the more the words underestimated by LSA correspond to that topic.

4.1.2 LSA - Lexical Chains Comparison

Using the LSA similarity between words, we computed a score ranging from 0 to 1 for every lexical chain.

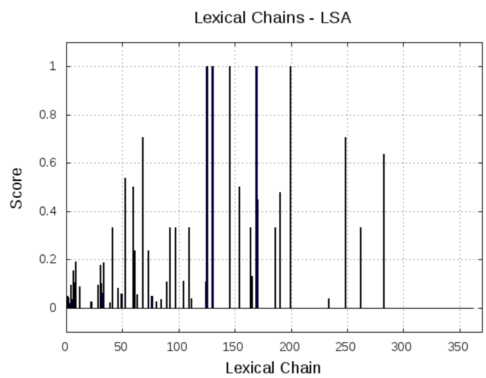

Figure 3: LSA scores for the lexical chains of the *tennis* article from the Wikipedia corpus.

For an example of the obtained results, see Figure 3 (for Topic 1 - *tennis*) where on the *x*-axis are the lexical chains (excluding those formed only by one word) and on the *y*-axis are their LSA scores.

We have noticed that the best lexical chains are obtained for the texts that had also a good correlation between the scores obtained by LDA and LSA. Also, one can see that there are only few lexical chains which are closely related in terms of LSA, which leads us to believe that LSA and lexical chains are not very well correlated.

Approximately 70% of the generated lexical chains were composed of a single word. In the rest of the lexical chains, the most frequent ones are those having small LSA scores - in the range (0, 0.25]. The other intervals represent only a small percent from the number of chains remaining when the single word chains are ignored.

The LSA scores are dependent on the lexical chain length, so we considered that it would be interesting to draw a parallel between these two elements. In Figure 4 are plotted the lexical chains lengths with their corresponding LSA scores for the tennis article. The *x*-axis contains the lexical chains indexes and the *z*-axis contains the LSA score and the length of that chain.

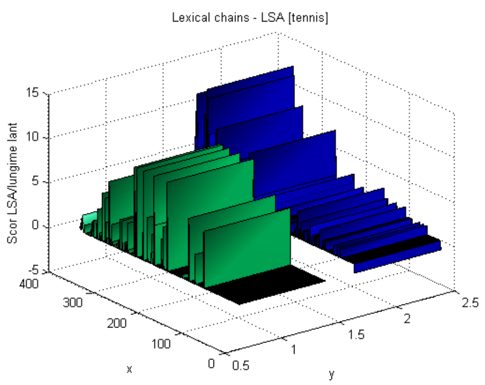

Figure 4: The LSA scores (green) and the lexical chains length (blue) from the *tennis* article.

4.1.3 LDA - Lexical Chains Comparison

For this comparison, we generated the most representative words for each of the four topics keeping the top 35, 50, 100 and 200 words and gradually comparing the number of common words between the topics and the lexical chains. It should be mentioned that a word can be representative for multiple topics (having different probabilities for each topic). The maximum lengths of the lexical chains from each article were 31, 28, 24 and 12 words for the articles about volcanoes, graffiti, astrology and tennis respectively. In the case of LDA topics having 35 words, the common words between LDA and lexical chains were:

- *Volcano* article: volcano, lava, surface, example, extinct, flow, explosive, water, generally, volcanism, fire, form, fluid, field, few, weight, first;
- *Tennis* article: tennis, game, player, first, court, point, french, receiver, real, playing, wide, cup, usually, full, current, covered, recent;
- *Graffiti* article: graffiti, new, culture, form, york, design, popular, hip, style, spray, paint, early, different, day, rock, history, elements, stencil, due, chicago, dragon, disagreement, newspaper, egypt, popularity, production;
- *Astrology* article: astrology, chinese, personality, scientific, birth, belief, challenge, astronomical, astronomy, avoid, philosophy, babylonian, basis, basic, average, birthday, beginning, century, believe.

In order to compare the results between LDA and lexical chains, we determined how many chains contained words that were also considered representative for the four LDA topics along with the number of such common words.

First of all, we computed for each topic the first 35 words and represented the frequency of common words between the lexical chains and the topics of this size. In this case, most chains had no common words with any of the topics (more than 700 such chains). The Topic 0 (*graffiti*) had one common word with the largest number of lexical chains (over 25 chains), the Topic 1 (*tennis*) had a common word with 17 such chains, while the last topic (*volcano*) had words in 15 lexical chains. Topic 2 (*astrology*) had two common words with 3 lexical chains (most chains comparing with the other topics), but had a smaller number of lexical chains (13) with which it had a single word in common. As an overall statistic, the words from Topic 0 (*graffiti*) could be found in the most lexical chains. After we increased the number of words to 50 per topic, around 430 chains had no word in common with the topics, and the number of most common words between topics and lexical chains increased to 3, although there were only two such chains – one for Topic 1 (*tennis*) and one for Topic 3 (*volcano*). Further increasing the number of words in a topic to 100, we saw that Topic 3 (*volcano*) had 4 common words with one lexical chain and, compared to the previous case, this time all the topics have found 3 common words with at least one lexical chains. At this point, Topic 1 (*tennis*) had a single word in common with over 40 lexical chain, this becoming the best score, comparing to the previous cases when the Topic 0 (*graffiti*) was the most popular in this category. Overall, the Topic 3's words are the most often found in the lexical chains (over 40 chains having o word in common, 2 having 2 words in common and 1 with 3 and 1 with 4 words in common).

Finally we increased the number of words per topic to 200 (Figure 5). Also in this case, there still remained around 350 chains that had no words in common with any of the topics. It can be seen that the Topic 3 (*volcano*) has 7 words in common with one of the lexical chains (the best score so far), while Topic 2 (*astrology*) had 5 common words with one of the chains. The details of this discussion are summarized in Table 4.

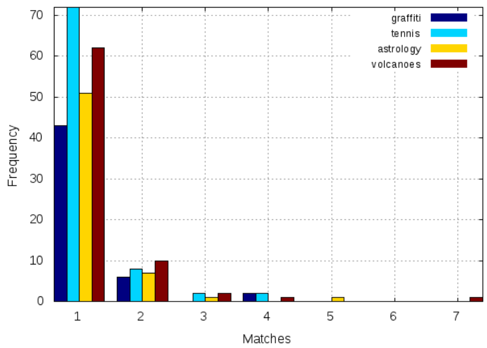

Figure 5: The distribution of the common words between topics (of 200 words) and the lexical chains.

Table 4: Number of chains having a single word in common with different topics (highest values are in bold), and the maximum number of words in common with a topic in a single chain.

Topic words/ topic	T0	T1	T2	T3	No topic	Max. common words
35	**>25**	16	12	15	>300	2 (3 chains for T2, 1 for the rest)
50	**29**	17	15	20	~300	3 (T1 & T3)
100	24	**>40**	33	>40	~270	4 (T3)
150	34	51	41	**>50**	~260	6 (T3)
200	>40	**>70**	>50	>60	~250	7 (T3)

In conclusion, the most frequent situation (besides the lexical chains having no word in common with the topics) is the one when the lexical chains and the topics have exactly one common word, and the maximum number of common words that was found was 7 for topics consisting of 200 words.

4.2 Chat Conversations Corpus

A similar methodology was used to compare the results on the chat corpus in order to see if there are any noticeable differences due to the change of the type of discourse. The results are reported more briefly in this section.

4.2.1 LDA - LSA Comparison

Table 5 presents the top 10 words from the 5 LDA topics. In Table 6 we present the most similar 30 word-pairs generated by LSA.

Similarly to the Wikipedia corpus, we plotted the distributions of LDA and LSA scores for each word from that topic and obtained the best result for Topic 1 (0.73). This case is presented in Figure 6, while in Figure 7 we present the scattering diagram for this topic. The Pearson's and the Spearman's Rank correlation coefficient between the LDA and LSA scores for each LDA topics are presented in Table 7.

Table 5: Top 10 words from the LDA topics in the chat corpus.

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Forums	wiki	blogs	chat	blog
Internet	solutions	brain storming	information	person
Good	solve	company	friends	forum
Ideas	opinion	clients	find	board
Right	web	changes	folksonomy	certain
Users	wave	compare	follow	fun
write	number	cases	great	new
idea	need	different	hard	part
people	like	easy	integrate	change
help	use	more	maybe	friend

Table 6: Top 30 most similar word-pairs generated by LSA in the chat corpus.

LSA Word Pairs		
traveller-messaging	patterns-vmtstudents	mathematicians-patterns
sets-colinear	flame-wars	dictate-behaviour
decides-improper	physically-online	satisfaction-conducted
easely-switchboard	inconvenient-counterargument	counts-popularity
ads-revenue	induction-patterns	editors-objectivity
supplying-focuses	inconvenient-counter	duties-minimum
patient-recall	sets-colinear	decides-improper
hm-conversions	equations-quicksilver	lie-proud
secure-hacked	simplifies-equals	chatroom-leaves
careful-posible	fellow-worker	hexagonal-array

As it was expected, the results for the chat corpus are less correlated than the ones obtained for the Wikipedia corpus. This drop in performance can be partly explained by the increased number of topics (one additional topic), but mostly by the different nature of discourse: the Wikipedia articles are much

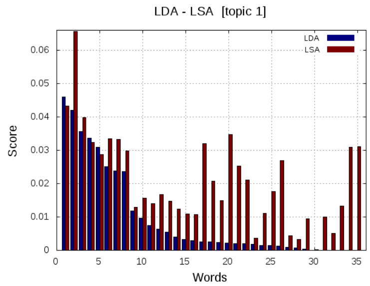

Figure 6: LDA – LSA distributions for Topic 1 from the chat corpus.

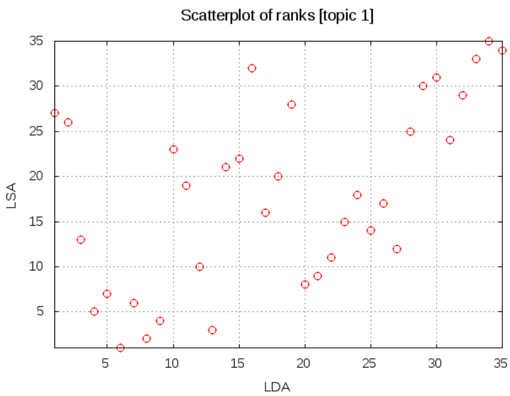

Figure 7: Scattering plot for the ranks distributions for the LDA–LSA comparison for Topic 1 from the chat corpus.

Table 7: LDA-LSA Pearson's Coefficient for the chat corpus.

Topic	Pearson's Coefficient	Spearman's Coefficient
0	0.63	0.46
1	**0.73**	**0.55**
2	0.55	0.41
3	0.46	0.35
4	0.71	0.32

more focused/cohesive and coherent than chat conversation between multiple participants. It also provides an insight related to the content of the chat conversations: it seemed that the topic one (related to *wikis/Wikipedia*) discovered by LDA was more coherent than the other topics, at least by looking at the LSA correlation scores. The second highest score in this hierarchy was for the *forum-blog* topic showing that the participants do not perceive significant differences between these concepts. However, the most intriguing result was the placement of the third topic (related to *chat*) on the last place, showing the least coherence. We expected that this topic to have in fact the highest coherence, being the tool most frequently used by the participants and therefore the tool that they knew best. These results may also be influenced by the way we are measuring the coherence of a LDA topic through its correlation with the average LSA similarity scores.

4.2.2 LSA - Lexical Chains Comparison

For the chat corpus, the values of the LSA similarity between words for every lexical chain ranged from -1 to 1, as it can be seen in Figure 8. We can observe that the correlation between the LSA and lexical chains for the chat corpus is lower than the one for the Wikipedia corpus, this fact being generated by the lower cohesion of the text in this case.

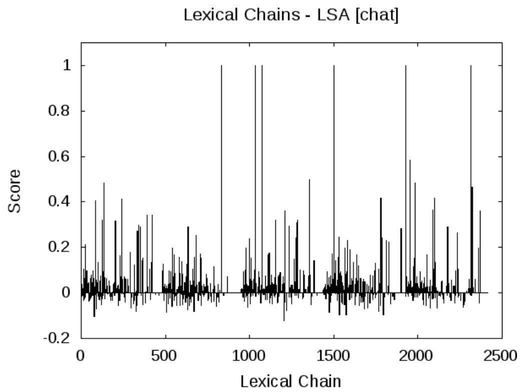

Figure 8: LSA scores for the lexical chains from the chat corpus.

4.2.3 LDA - Lexical Chains Comparison

Similarly to the Wikipedia corpus, each of the five topics was generated keeping the top 35, 50, 100 and 200 words and gradually comparing the number of common words between the topics and the lexical chains. The maximum length of the lexical chains from this corpus was 84, much larger than the one obtained in the case of the Wikipedia corpus. This is due to the fact that the four topics imposed for debating in the chat conversations (forum, chat, blog, and wikipedia) were strongly related compared to the Wikipedia articles that debated topics from different domains.

The number of common words is predominantly 1, reaching a maximum of 8 common words for the third topic (related to chat) for a length of the lexical chain of 150 words. The results are similar to those obtained for the Wikipedia corpus.

5 CONCLUSIONS

In this paper we discussed the characteristics and behaviour of three methods frequently used to assess semantics in various NLP applications: LSA, LDA and lexical chaining. These methods have been tested on two different corpora containg different types of written discouse: a corpus consisting of 4 articles from Wikipedia and another one consisting of 35 chat conversations with multiple participants debating four pre-imposed topics: forum, chat, blog and wikis.

In contrast with the previous studies, we have compared the outcomes of the three methods using quantitative scores computed based on the outputs of each method. These scores included correlations between similarity scores and the number of common words from topics and chains. Thus, the obtained results are task and discourse-independent.

The most important result is that LSA and LDA have shown the strongest correlation on both corpora. This is consistent with the theoretical underpinnings, as LDA is similar to Probabilistic Latent Semantic Analysis (pLSA), except that the LDA distribution of topics is assumed to have a prior Dirichlet distribution. Moreover, LSA scores might be used to compute the coherence of a LDA topic as shown in the paper.

Another important contribution is that WordNet-based lexical chains are not very correlated with neither LSA nor LDA, therefore they might be seen as complementary to the LSA or LDA results.

ACKNOWLEDGEMENTS

This research was supported by project No.264207, ERRIC-Empowering Romanian Research on Intelligent Information Technologies/FP7-REGPOT-2010-1.

REFERENCES

Barzilay, R. and Elhadad. M., 1997. Using lexical chains for text summarization. In: *Proceedings of the Intelligent Scalable Text Summarization Workshop*, pp. 10–17.

Budanitsky, A. and Hirst, G., 2006. Evaluating wordnet-based measures of semantic relatedness. In: *Computational Linguistics 32 (1)*, pp. 13–47.

Blei, D. M., Ng, A. Y. and Jordan, M. I., 2003. Latent Dirichlet allocation. In: *Journal of Machine Learning Research 3*, pp. 993-1022.

Boyd-Graber, J., Fellbaum, C., Osherson, D. and Schapire, R., 2006. Adding dense, weighted, connections to WordNet. In: *Proceedings of the 3rd GlobalWordNet Meeting*, pp. 29–35.

Carthy, J., 2004. Lexical chains versus keywords for topic tracking. In: *Computational Linguistics and Intelligent Text Processing, LNCS*, pp. 507–510. Springer.

Chiru, C., Janca, A., Rebedea, T., 2010. Disambiguation and Lexical Chains Construction Using WordNet. In S. Trăușan-Matu, P.Dessus (Eds.) Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity, MatrixRom, pp 65-71.

Cramer, I., 2008. How well do semantic relatedness measures perform? a meta-study. In: *Proceedings of the Symposium on Semantics in Systems for Text Processing*.

Griffiths, T. L., Steyvers, M. and Tenenbaum, J. B., 2007. Topics in semantic representation. In: *Psychological Review*, vol. 114, no. 2, pp. 211–244.

Gong, Y. and Liu, X., 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In: *Proceedings of the 24th ACM SIGIR conference*, pp. 19-25.

Haghighi, A. and Vanderwende, L., 2009. Exploring content models for multi-document summarization. In: *Proceedings of HLT-NAACL*, pp. 362–370.

Halliday, M. A.K. and Hasan, R., 1976. *Cohesion In English*, Longman.

Jiang, J. J. and Conrath, D. W, 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: *Proceedings of ROCLING X*, pp. 19-33.

Kakkonen, T., Myller, N., Sutinen, E.and Timonen, J., 2008. Comparison of Dimension Reduction Methods for Automated Essay Grading. In: *Educational Technology & Society*, 11(3), pp. 275–288.

Landauer, T. K. and Dumais, S. T., 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2), 211-240.

Misra, H., Yvon, F., Jose, J. and Cappé, O., 2009. Text Segmentation via Topic Modeling: An Analytical Study. In: *18th ACM Conference on Information and Knowledge Management*, pp. 1553–1556.

Morris, J. and Hirst, G., 1991. Lexical Cohesion, the Thesaurus, and the Structure of Text. In: *Computational Linguistics*, Vol 17(1), pp. 211-232.

Novischi, A. and Moldovan, D., 2006. Question answering with lexical chains propagating verb arguments. In: *Proceedings of the 21st International Conference on CL and 44th Annual Meeting of ACL*, pp. 897–904.

Tsatsaronis, G., Varlamis, I. and Vazirgiannis, M., 2010. Text relatedness based on a word thesaurus. In: *Artificial Intelligence Research*, 37, pp. 1–39.