Time and Semantic Similarity – What is the Best Alternative to Capture Implicit Links in CSCL Conversations?

Gabriel Gutu, Mihai Dascalu, Traian Rebedea, and Stefan Trausan-Matu,
gabriel.gutu@cs.pub.ro, mihai.dascalu@cs.pub.ro, traian.rebedea@cs.pub.ro, stefan.trausan@cs.pub.ro
University Politehnica of Bucharest, Computer Science Department

Abstract: The goal of our research is to compare novel semantic techniques for identifying implicit links between utterances in multi-participant CSCL chat conversations. Cohesion, reflected by the strength of the semantic relations behind the automatically identified links, is assessed using WordNet-based semantic distances, as well as unsupervised semantic models, i.e. Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). The analysis is built on top of the *ReaderBench* framework and multiple identification heuristics were compared, including: semantic cohesion metrics, normalized cohesion measures and Mihalcea's formula. A corpus of 55 conversations in which participants used explicit links between utterances where they considered necessary for clarity was used for validation. Our study represents an in-depth analysis of multiple methods used to identify implicit links and reveals the accuracy of each technique in terms of capturing the explicit references made by users. Statistical similarity measures ensured the best overall identification accuracy when using Mihalcea's formula, while WordNet-based techniques provided best results for un-normalized similarity scores applied on a window of 5 utterances and a time frame of 1 minute.

Introduction

Chat represents a commonly used collaboration tool nowadays that can also be successfully employed in learning processes, such as Computer Supported Collaborative Leaning (CSCL) (Stahl, 2006). Creativity fostering is a key element in multi-participant chat conversations (Trausan-Matu, 2010), where multiple changes of perspectives and of points of interest are frequently encountered, which are helpful for social knowledge building in educational settings. However, these mixed discussion topics and threads may be difficult to follow and understand and thus, participants add explicit links to previous utterances when they have this facility (Holmer, Kienle, & Wessner, 2006) in order to ensure threading and, consequently, coherence between utterances. As this is a cumbersome task for many users, which tend to introduce few explicit links (if any), the need of automation has dramatically increased when analyzing chats of hundreds of utterances and with more than 3 participants. This process of linking related utterances is referred to as *implicit links detection* (Trausan-Matu & Rebedea, 2010) and represents an important step that allows the integration of additional operations on texts such as topic mining, sentiment analysis, detection of lexical chains, and evaluating the degree of collaboration in problem solving and CSCL.

Natural Language Processing (NLP) techniques (Manning & Schütze, 1999) are more and more used nowadays since they provide efficient analyses of written texts. In contrast to other types of web collaboration tools such as forums or social networks, most chat systems do not provide any "*reply-to*" option. This lack makes difficult to follow the threads of discussions in chats with more than two participants, generating discourse segmentation. Therefore, due to this lack of a referencing facility in the clear majority of online chats, the usage of NLP tools for the detection of implicit links between utterances represents an important research topic.

The purpose of this paper is to determine which state of the art semantic similarity measure performs best for the detection of implicit links in multi-party chat conversations and what is the optimum distance in terms of utterances and time frames to look for them. The corpus for this comparative analysis consists of a collection of 55 conversations lasting up to two hours, performed by computer science students from our faculty using the *ConcertChat* environment (Holmer, Kienle, & Wessner, 2006), which enables users to explicitly reference previous utterances. Within these conversations, participants had to discuss about the benefits and disadvantages of each several web collaboration technologies (i.e., wiki, blog, forum chat) and identify the most suitable tool to be used by an enterprise (Trausan-Matu & Rebedea, 2010). This collection represents a refined version of an initial set of 200 conversations used in previous studies (Gutu, Rebedea, & Trausan-Matu, 2015) based on the following criteria: multiple conversation sessions with the same participants were discarded, as well as discussions with a limited timeframe (less than 30 minutes), too few utterances (less than 50 utterances) or too few explicit links added by users (less than 10 explicit links per conversation).

In terms of the structure of the paper, the following section provides general information about semantic similarity and introduces the five measures available in the *ReaderBench* framework (Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014; Dascalu et al., 2015a; Dascalu et al., 2015b): *Latent Semantic Analysis* (LSA)

(Landauer & Dumais, 1997), *Latent Dirichlet Allocation* (LDA) (Blei, Ng, & Jordan, 2003) and three *WordNet*-based distance functions: Leacock Chodorow (Leacock & Chodorow, 1998), Wu Palmer (Wu & Palmer, 1994) and path length (Budanitsky & Hirst, 2006). The third section presents the results of our analysis alongside statistical information about the chat corpus. In the end, conclusions and future work are presented.

Related work on semantic models

Semantic cohesion reflects the degree to which two text fragments are related one to another in terms of meaning (Bestgen, 2012) and can be automatically evaluated using several approaches. In previous studies in the Natural Language Processing field, several techniques gained high popularity. The first one consists of applying different semantic distance functions on ontologies (Budanitsky & Hirst, 2006), such as the WordNet lexical database (Miller, 1995). Second, Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) is the most frequently used method to compute semantic similarity by relying on vector spaces of keywords (terms). Third, probabilistic topic modeling has gained an increasing attention lately, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) being the most frequently used method of this kind.

LSA (Landauer & Dumais, 1997) is a NLP technique based on a vector space model highlighting term co-occurrences within documents. LSA is frequently used to compute the similarity between documents and between terms (Manning & Schütze, 1999). LSA is based on a "bag of words" approach in which word order is disregarded and words' occurrences are normalized through Term frequency – Inverse Document frequency or log entropy. A singular-value decomposition (SVD) (Golub & Reinsch, 1970), followed by a reduction of the matrices' dimensionality through a projection on k dimensions is performed in order to determine indirect links induced between groups of terms and underlying documents. The optimal empiric range for k is 300±50 (Landauer, McNamara, Dennis, & Kintsch, 2007; Lemaire, 2009). LSA can be perceived as a mathematical optimization for representing the meaning of words and group of words in a vector space by adopting an unsupervised learning technique applied on a corpus of texts. Our LSA model was trained on a pre-processed version of a custom corpus obtained by concatenating the TASA corpus (Touchstone Applied Science Associates, Inc., http://lsa.colorado.edu/spaces.html) that contains general texts, novels and newspaper articles and a corpus of more than 500 CSCL-related scientific papers. Stop-words and non-dictionary words were disregarded, inflectional word forms were reduced to their lemmas and only paragraphs with more than 20 content words were considered.

LDA (Blei, Ng, & Jordan, 2003) is a generative probabilistic process built on top of the assumption that documents integrate multiple topics and can be therefore considered a mixture of corpus-wide topics. Each topic represents a Dirichlet distribution (Kotz, Balakrishnan, & Johnson, 2000) over the vocabulary where related concepts have similar probabilities based on co-occurrence patterns from the training corpora. Although each topic contains all the words from the vocabulary, a clear differentiation in terms of corresponding probabilities can be observed between salient versus dominant concepts. Similar to LSA, LDA relies on the "bag of words" approach and classifies new texts in terms of the latent topics inferred from the model trained on a text collection. Documents and words alike become topics distributions drawn from Dirichlet distributions, while semantic similarities between textual fragments are determined using the Jensen-Shannon dissimilarity (JSH) (Manning & Schütze, 1999), a symmetric smoothed alternative of the KL divergence (Kullback & Leibler, 1951). A drawback of the traditional LDA model is that it uses an imposed number of topics. (Teh, Jordan, Beal, & Blei, 2006) have introduced an alternative, Hierarchical Dirichlet Process (HDP), a nonparametric Bayesian approach used to cluster grouped data. The HDP mixture model can be considered as a generalization of LDA in which the number of topics is inferred from the training text corpora (Teh, Jordan, Beal, & Blei, 2006). HDP applied on the custom-built TASA–CSCL corpus inferred 118 topics as the optimal number of topics.

WordNet (Miller, 1995) represents a lexical ontology having words organized in four different trees based on their corresponding parts of speech: nouns, verbs, adverbs or adjectives. The hierarchical representation of words using synsets (sets of synonyms) describes relations between them. Thus, a word which is a descendant of another concept in WordNet's internal representation is a more specific term, while the parent represents a more general concept. Given WordNet's internal representation, several distance measures were developed, out of which the most notable three, implemented in the *ReaderBench* framework, were used in this study: Leacock Chodorow (Leacock & Chodorow, 1998), Wu Palmer (Wu & Palmer, 1994) and path length (Budanitsky & Hirst, 2006), all presented in detail in Table 1.

Table 1: Semantic distances based on WordNet that are used in the study

Semantic distance	Formula	Description
Leacock Chodorow	$sim_{LC}(c_1, c_2) = -log\frac{l(c_1, c_2)}{2D}$	The path length normalized by the depth of the ontology, where depth represents the path length from the current concept to the global root.
Wu Palmer	$sim_{WP}(c_1, c_2) = \frac{2 \times d(lso(c_1, c_2))}{l(c_1, lso(c_1, c_2)) + l(c_2, lso(c_1, c_2)) + 2 \times d(lso(c_1, c_2))}$	Similarity is computed considering the depths of the two concepts, as well as their least common ancestor.
Path length	$l(c_1, c_2)$	The length of the shortest path between the two concepts.

Method

Two accuracy measures were considered within this study: a) *exact* implicit links detection, when the computed reference is the same as the explicit reference attribute set by the chat participant, and b) *in-turn* implicit links detection, when the computed reference belongs to the same turn (i.e., a collection of adjacent utterances belonging to the same participant, including the utterance mentioned within the explicit link). Our corpus of 55 chat conversations was initially cleaned using several NLP refinements (Manning & Schütze, 1999): *stop-words* (words with no semantic relevance and no contextual information) were eliminated, duplicate words frequently encountered in chat conversations were removed and the remaining words were lemmatized using the CoreNLP library (Manning et al., 2014).

Table 2 introduces two examples of identified implicit links extracted from the same conversation. The examples show the differences between in-turn matching, when implicit links must belong to the same participant in a continuous block of utterances, while in case of exact-matching implicit links must overlap perfectly with the explicit links defined by the user. The first example shows utterance with id 74 having an explicit reference to utterance 65 which was manually added by a participant (explicit links are in the third column – *Ref. ID*). However, when imposing a window of maximum 5 utterances and 1 minute time frame, the detected implicit link is utterance 72 (emphasized) using the Path Length similarity measure. As turn 72 is enclosed in a continuous series of utterances belonging to the same user (i.e., Monica), we consider this to be a correct *in-turn matching*, but an incorrect *exact matching*. In the second excerpt, the identified implicit link for utterance 140 is turn 138 using the same parameters for time frame, distance and semantic similarity. Turn 138 is also the explicit reference for utterance 140, as can be easily observed from the *Ref. ID* column, and this is a correct *exact matching*.

Table 2: Excerpt from a chat conversation

Measure	Utt. ID	Ref. ID	Speaker	Time	Content
In-turn matching	65		Monica	09:08:27	features to add RSS feeds, file sharing and so on
	… (several utterances of the same participant, Monica) ...				
	72		**Monica**	**09:09:57**	**and they embed only what you need**
	73		Monica	09:10:16	users tend to be scared away by a multitude of features that they need to figure out
	74	65	Razvan	09:10:22	The thing that I think would be a problem with wikis is that they will not allow a person to keep track of the latest information added. Ok RSS are good but not everybody wants to use an RSS feed reader.
	…				
Exact matching	135		Monica	09:21:18	actually apart from the user setup
	136		Monica	09:21:30	that personally I find quite an effort
	137		Monica	09:21:41	blogs are a good solution
	138		**Stefan**	**09:22:01**	**you know the biggest disadvantage of wikis? that anybody can input and that makes wikis a not-so-reliable source of info**
	139	138	Alex	09:22:20	that depends on the configuration
	140	138	Razvan	09:22:24	you could have admins that check the information

Results

Several measurements centered on per-chat analyses and distance-related or time-related statistical tests were performed on the corpus of chat conversations before running the implicit links detection process. The conversation corpus totals 17,612 utterances ($M = 320.22$, $SD = 136.04$, $min = 79$, $max = 817$) with an average of 4.35 participants per conversation ($SD = 0.97$, $min = 3$, $max = 8$) and 4,463 explicit references in total ($M = 81.15$, $SD = 45.29$, $min = 15$, $max = 226$, per chat). The average coverage (i.e., the percentage of referred utterances from the total number of utterances from a chat) was 28.62% ($SD = 16.62\%$, $min = 5\%$, $max = 72.9\%$,). The average time duration of a conversation is about 2 hours.

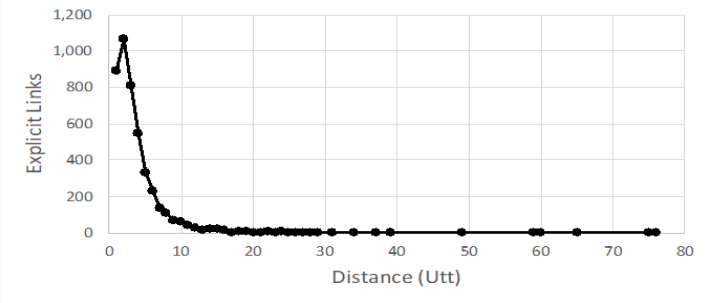


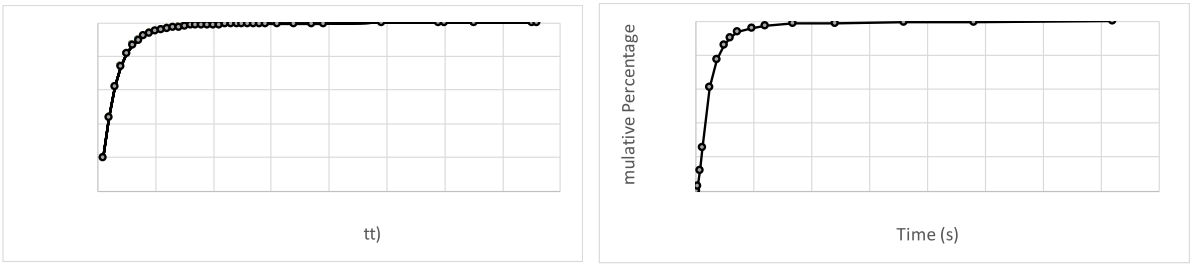
Figure 1. Distribution of explicit links per distance between turns.

Table 2 presents the statistical data from which it can be easily observed that the number of references decreases exponentially with distance (see Figure 1 for a visual representation of the distances' distribution for the entire corpus of conversations). Moreover, distances (D) of 1 to 10 cover more than 95% of the references, while a window of 20 utterances covers more than 99% of potential explicit links. In terms of authors of the original and referred utterance, as expected, there is a higher percentage of explicitly linked utterances belonging to different speakers (M = 88.51%) than to the same chat participant for a window size of 20 adjacent utterances.

Figure 2 presents the graphical evolution of the coverage of explicit links as a function of distance and time. We can observe in a visual manner that a window size of 20 utterances ensures the coverage of most (99%) explicit links, while a distance of 10 utterances enables a sufficient level of certainty (covering more than 95% of the total links). Given these distributions, we decided to compute the semantic similarity between each utterance and the previous ones considering window sizes of 20, 10 and 5 utterances respectively.

Table 3: Explicit links occurrences and coverage in terms of distance

Distance	Explicit links	M (SD)	Min / Max	Percentage		Same speaker		Different speaker	
				Local	Cumulative	#	%	#	%
1	890	16.18 (13.10)	0 / 54	19.94%	19.94%	213	23.93%	677	76.07%
2	1065	19.36 (11.92)	1 / 49	23.86%	43.80%	126	11.83%	939	88.17%
3	810	14.73 (9.84)	2 / 42	18.15%	61.95%	79	9.75%	731	90.25%
4	548	9.96 (6.56)	1 / 31	12.28%	74.23%	84	15.33%	464	84.67%
5	332	6.04 (4.61)	0 / 24	7.44%	81.67%	44	13.25%	288	86.75%
6	230	4.18 (3.36)	0 / 16	5.15%	86.83%	26	11.30%	204	88.70%
7	134	2.44 (2.39)	0 / 11	3.00%	89.83%	10	7.46%	124	92.54%
8	106	1.93 (1.69)	0 / 7	2.38%	92.20%	6	5.66%	100	94.34%
9	68	1.24 (1.39)	0 / 6	1.52%	93.73%	5	7.35%	63	92.65%
10	62	1.13 (1.33)	0 / 6	1.39%	95.12%	7	11.29%	55	88.71%
11	43	0.78 (1.03)	0 / 4	0.96%	96.08%	6	13.95%	37	86.05%
12	30	0.55 (0.83)	0 / 3	0.67%	96.75%	6	20.00%	24	80.00%
13	16	0.29 (0.60)	0 / 2	0.36%	97.11%	2	12.50%	14	87.50%
14	20	0.36 (0.36)	0 / 2	0.45%	97.56%	3	15.00%	17	85.00%
15	21	0.38 (0.71)	0 / 3	0.47%	98.03%	1	4.76%	20	95.24%
16	17	0.31 (0.57)	0 / 2	0.38%	98.41%	1	5.88%	16	94.12%
17	4	0.07 (0.33)	0 / 2	0.09%	98.50%	0	0.00%	4	100.00%
18	9	0.16 (0.46)	0 / 2	0.20%	98.70%	3	33.33%	6	66.67%
19	6	0.11 (0.37)	0 / 2	0.13%	98.83%	1	16.67%	5	83.33%
20	4	0.07 (0.26)	0 / 1	0.09%	98.92%	1	25.00%	3	75.00%



(a) (b)

Figure 2. Cumulative coverage of explicit links (a) per distance and (b) per time.

Regarding the time difference between utterances, several time frames were set and explicit links' coverage within these time frames were computed. Table 3 shows the coverage of the links within the selected time frames. Significant changes in terms of cumulative percentage were desired, which made us to select 5 time frames for the study: 30 seconds, 1, 2, 3 and 5 minutes. As it was observed, within 5 minutes 97% of the links are covered, while the time frame of 1 minute covers 61% of them.

Implicit links identification

Three metrics were used in order to identify the best semantic match between the current utterance and all previous ones within the imposed window sizes: semantic similarity (SIM), normalized similarity by inverse distance between current utterance and referred utterance (NSIM), and Mihalcea's similarity formula (MSIM) (Mihalcea, Corley, & Strapparava, 2006). SIM represents the baseline similarity metric for each semantic model; for example, when talking about LSA, we use SIM to refer to the standard formula for computing LSA cosine similarity. NSIM is used to refer to a normalized value of the previously introduced similarity. The specific formula developed by Mihalcea was introduced as a third metric named here MSIM. In a nutshell, SIM, NSIM and MSIM refer to specific optimization formulas which rely on the underlying semantic models.

The implicit link reflects the highest score within the window in terms of distance or time frame, namely the pair of utterances having the highest semantic similarity between them. Accuracy was assessed in terms of the overlap between the automatically identified implicit links and the explicit ones added by users on two criteria: *exact matching* and *in-turn matching*, presented in pairs delimited by slash in all subsequent tables. Table 5 presents the percentage of detected explicit links using both exact and in-turn accuracy measures per window size. Table 6 presents the same percentages computed per time frame. Emphasized values represent the semantic similarity formula that provided the best accuracy for each technique and for each windows size.

Table 4: Explicit links occurrences and coverage in terms of time frame

Time	Explicit links	Cumulative Percentage	Same speaker		Different speaker	
			#	%	#	%
1 sec	1	0.03%	0	0.00%	1	100.00%
2 secs	1	0.03%	0	0.00%	1	100.00%
3 secs	2	0.05%	0	0.00%	2	100.00%
5 secs	4	0.11%	1	25.00%	3	75.00%
10 secs	100	2.71%	39	39.00%	61	61.00%
20 secs	459	12.42%	126	27.45%	333	72.55%
30 secs	957	25.89%	222	23.20%	735	76.80%
1 min	2246	60.77%	364	16.21%	1882	83.79%
1.5 mins	2850	77.11%	423	14.84%	2427	85.16%
2 mins	3176	85.93%	453	14.26%	2723	85.74%
2.5 mins	3342	90.42%	467	13.97%	2875	86.03%
3 mins	3451	93.37%	477	13.82%	2974	86.18%
4 mins	3553	96.13%	487	13.71%	3066	86.29%
5 mins	3593	97.21%	493	13.72%	3100	86.28%
7 mins	3634	98.32%	496	13.65%	3138	86.35%
10 mins	3658	98.97%	499	13.64%	3159	86.36%
15 mins	3673	99.38%	503	13.69%	3170	86.31%
20 mins	3682	99.62%	505	13.72%	3177	86.28%
30 mins	3685	99.70%	505	13.70%	3180	86.30%
1 hour	3688	99.78%	507	13.75%	3181	86.25%
1.5 hours	3689	99.81%	507	13.74%	3182	86.26%
2 hours	3690	99.84%	507	13.74%	3183	86.26%
> 2 hours	3696	100.00%	508	13.74%	3188	86.26%

Table 5: Percentage of correctly matched explicit links per window size (Exact matching / In-turn matching)

Window size	Measure	LSA	LDA	Leacock	Wu-Palmer	Path length
20 (chance = 5%)	SIM	17.47% / 22.99%	14.84% / 20.48 %	15.46% / 21.55%	17.99% / 24.11%	20.27% / 26.6 %
	NSIM	**29.27% / 37.98%**	**26.48% / 35.88%**	**26.40% / 35.82%**	**27.48% / 36.72%**	**31.00% / 39.98%**
	MSIM	17.97% / 23.47%	19.02% / 24.35%	14.03% / 19.40%	16.67% / 22.10%	20.05% / 26.54%
10 (chance = 10%)	SIM	23.41% / 31.19%	20.18% / 28.30%	21.67% / 30.03%	24.69% / 32.81%	26.62% / 34.99%
	NSIM	**29.82% / 38.65%**	**26.53% / 35.88%**	**27.03% / 36.44%**	**28.08% / 37.31%**	**31.45% / 40.48%**
	MSIM	24.03% / 31.36%	24.93% / 32.76%	19.22% / 27.31%	21.86% / 29.67%	25.00% / 32.80%
5 (chance = 20%)	SIM	29.23% / 38.79%	27.31% / 37.84%	**27.65% / 37.49%**	**30.19% / 39.87%**	31.66% / 41.03%
	NSIM	**30.50%** / 39.49%	26.35% / 35.72%	27.29% / 36.89%	28.31% / 37.62%	**31.99% / 41.29%**
	MSIM	30.16% / **39.51%**	**30.08% / 39.82%**	25.76% / 35.97%	27.88% / 38.08%	30.05% / 39.43%

Table 6: Percentage of correctly matched explicit links per time frame (Exact matching / In-turn matching)

Time frame	Measure	LSA	LDA	Leacock	Wu-Palmer	Path length
5 mins	SIM	20.31% / 27.28%	16.85% / 23.70%	18.21% / 25.46%	20.53% / 27.40%	22.51% / 29.62%
	NSIM	**29.22% / 37.91%**	**26.28% / 35.58%**	**26.19% / 35.46%**	**27.27% / 36.39%**	**30.85% / 39.71%**
	MSIM	20.50% / 27.03%	21.85% / 28.57%	16.40% / 23.59%	18.92% / 25.85%	22.03% / 29.06%
3 mins	SIM	25.22% / 33.50%	22.42% / 30.69%	24.15% / 32.86%	26.41% / 34.81%	28.12% / 36.49%
	NSIM	**29.98% / 38.74%**	26.29% / **35.55%**	**26.95% / 36.41%**	**27.94% / 37.12%**	**31.71% / 40.82%**
	MSIM	25.81% / 33.60%	**26.78%** / 34.76%	21.79% / 30.63%	24.15% / 32.74%	27.18% / 35.19%
2 mins	SIM	29.03% / 37.38%	26.75% / 35.95%	**27.52% / 36.95%**	**30.10% / 39.45%**	31.49% / 40.41%
	NSIM	**30.56% / 39.23%**	25.55% / 34.83%	26.85% / 36.24%	27.86% / 36.94%	**31.70% / 40.79%**
	MSIM	29.82% / 37.91%	**29.52% / 38.13%**	25.79% / 35.11%	27.69% / 36.61%	30.27% / 38.82%
1 min	SIM	28.74% / 37.20%	27.11% / 35.98%	**27.97% / 36.95%**	**29.12% / 37.96%**	**29.62% / 38.46%**
	NSIM	27.17% / 35.73%	22.87% / 31.75%	24.72% / 33.51%	25.41% / 34.07%	28.20% / 36.72%
	MSIM	**28.87% / 37.25%**	**28.78% / 37.15%**	26.89% / 35.71%	27.68% / 36.55%	29.04% / 37.73%
30 secs	SIM	**15.13% / 22.97%**	14.98% / 22.94%	**14.44% / 22.59%**	**14.48% / 22.62%**	**14.58% / 22.64%**
	NSIM	14.53% / 22.31%	13.20% / **20.78%**	13.67% / 21.58%	13.73% / 21.62%	14.11% / 22.12%
	MSIM	14.71% / 22.56%	**14.98%** / 22.93%	14.23% / 22.26%	14.25% / 22.36%	14.39% / 22.51%

All semantic measures per window size exceeded random chance, i.e. the percentage of matches by randomly selecting a reference from the imposed window. Normalized semantic similarity provided the best accuracy for most of the used techniques and for all the three selected window sizes: about 30% for perfect match and about 40% for in-turn match. As presented in Table 3, the window size of 10 utterances covers more than 95% of the explicit links; corroborated with a random chance of only 10%, NSIM produced significant results with almost 30% exact accuracy and 40% in-turn accuracy. Accuracies are just a little higher (less than 1%) for the window size of 20 which covers 99% of the explicit links, but with the disadvantage of doubling the computational effort. Thus, we conclude that the window size of 10 using the NSIM technique is the best alternative to be used for determining implicit links in chat conversations in terms of distance. The window size of 5 utterances provided close results, too. Both window sizes were considered for the further experiment.

In terms of time frames, the best results were obtained for the window size of 2 minutes, both for exact matching and in-turn matching. The 1-minute window provided close values to the 2-minutes window, hence both of the time frames were considered. Results for the subsequent analysis with window sizes of 5 and 10 utterances, and time frames of 1, respectively 2 minutes, are presented in Table 7. With regards to the optimal window to search for implicit links, for our corpus, a distance of 5 utterances and a time window of 1 minute provided the leading scores for both exact matching and in-turn matching. Furthermore, compared to previous results in Table 5 and Table 6, better similarity scores were obtained by narrowing both the distance and the time frame.

By analyzing the semantic similarity scores obtained for the chosen pairs of window sizes and time frames, the results show specificity based on the employed measure. Hence, using LSA, for a window of 10 utterances, the NSIM formula detects most links, while for a window of 5 utterances the MSIM formula is the best. When it comes to LDA, the MSIM formula detects most explicit links. The WordNet-based techniques provide the best results with the SIM formula. In terms of formula to be used, we can observe that WordNet-based techniques brought better results using the un-normalized semantic similarity, while for the statistical methods, namely LSA and LDA, Mihalcea's formula can be considered the best for most scenarios.

Table 7: Percentage of correctly matched explicit links per chosen (window size, time frame) pairs (Exact matching / In-turn matching)

Window size, time frame	Measure	LSA	LDA	Leacock	Wu-Palmer	Path length
5 utt, 1 min	SIM	30.68% / 39.72%	28.77% / 38.61%	**28.55% / 38.18%**	**30.95% / 40.44%**	**32.44% / 41.49%**
	NSIM	30.01% / 38.74%	25.43% / 34.73%	26.78% / 36.19%	27.73% / 36.81%	31.35% / 40.51%
	MSIM	**31.45% / 40.41%**	**30.98% / 40.09%**	27.06% / 36.75%	28.89% / 38.45%	31.04% / 39.97%
10 utt, 1 min	SIM	29.25% / 37.80%	27.12% / 36.27%	**27.82% / 37.30%**	**30.27% / 39.62%**	**31.88% / 40.78%**
	NSIM	**30.28% / 38.94%**	25.55% / 34.83%	26.91% / 36.29%	27.88% / 36.94%	31.57% / 40.65%
	MSIM	30.28% / 38.48%	**30.08% / 38.67%**	26.16% / 35.59%	27.99% / 36.95%	30.60% / 39.16%
5 utt, 2 mins	SIM	30.68% / 39.72%	28.77% / 38.61%	**28.55% / 38.18%**	**30.95% / 40.44%**	**32.44% / 41.49%**
	NSIM	30.01% / 38.74%	25.43% / 34.73%	26.78% / 36.19%	27.73% / 36.81%	31.35% / 40.51%
	MSIM	**31.45% / 40.41%**	**30.98% / 40.09%**	27.06% / 36.75%	28.89% / 38.45%	31.04% / 39.97%
10 utt, 2 mins	SIM	29.25% / 37.80%	27.12% / 36.27%	**27.82% / 37.30%**	**30.27% / 39.62%**	**31.88% / 40.78%**
	NSIM	30.28% / **38.94%**	25.55% / 34.83%	26.91% / 36.29%	27.88% / 36.94%	31.57% / 40.65%
	MSIM	**30.28%** / 38.48%	**30.08% / 38.67%**	26.16% / 35.59%	27.99% / 36.95%	30.60% / 39.16%

Conclusions and future work

In this research, our aim was to compare various methods for the identification of implicit links by computing semantic similarity between every utterance and the previous utterances, within imposed distance and time windows. The performed comparative analysis provides evidence that using a combined window with the previous 5 utterances and 1-minute pair provides the best trade-off in terms of both exact and in-turn accuracies for the implicit links detection process. Plus, WordNet-based techniques provide the best overall results if un-normalized semantic distances are used, while statistical techniques such as LSA and LDA provide best results when Mihalcea's formula is employed. Current results surpass previous studies that used only un-normalized semantic similarity measures (Rebedea, Chiru, & Gutu, 2014; Rebedea & Gutu, 2013) and provide a deeper analysis on how to achieve the best accuracy for identifying implicit links in chat conversations. Chat conversations represent a modern tool used for collaborative working and are widely used in educational environments and learning tasks as they allow automatic analyses of natural language such as topic mining, sentiment analysis, lexical chains and others. Detection of implicit links established between utterances represents an initial step in processes that refer to text cohesion in general. The results presented in this paper are intended to facilitate the development of a more advanced chat tool that can automatically extract such NLP-related information.

Practical outcomes of implicit links identification include the detection of students' interaction patterns in ongoing conversations, as well as the extraction of lexical chains and of discussion threads. Integrated within a chat client or any CSCL environment housing multi-participant conversations, these features may provide better awareness to users in terms of understanding and following the current discussion threads. Moreover, the identified implicit links facilitate tutor assessment while highlighting members' active involvement, as well as their collaboration with other participants.

However, several adjustments are considered for increasing the accuracy of the implicit links identification process. First, machine learning techniques could be used to create an aggregated similarity score relying on multiple semantic measures. Second, dynamic sliding windows could be enforced by considering cut-offs induced by topic changes or long pauses within the discourse. Third, certain patterns extracted using speech acts (e.g., continuations, question answering) (Searle, 1969) and discourse connectors may be indicative of implicit links within the discourse. For example, the presence of a contrast connective ("but") or of a question answering sequence may indicate an implicit link between two utterances that might have a low in-between semantic similarity score.

References

Bestgen, Y. (2012). Évaluation automatique de textes et cohésion lexicale. *Discours, 11*. doi: 10.4000/discours.8724

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics, 32*(1), 13–47.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learner productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (pp. 345–377). Cham, Switzerland: Springer.

Dascalu, M., Stavarache, L. L., Dessus, P., Trausan-Matu, S., McNamara, D. S., & Bianco, M. (2015a). ReaderBench: An Integrated Cohesion-Centered Framework. In G. Conole, T. Klobucar, C. Rensing, J. Konert & É. Lavoué (Eds.), *EC-TEL 2015* (pp. 505–508). Toledo, Spain: Springer.

Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P., Bianco, M., & McNamara, D. S. (2015b). ReaderBench: An Integrated Tool Supporting both Individual and Collaborative Learning. In *5th Int. Learning Analytics & Knowledge Conf. (LAK'15)* (pp. 436-437). Poughkeepsie, NY: ACM.

Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik, 14*(5), 403–420.

Gutu, G., Rebedea, T., & Trausan-Matu, S. (2015). A comparison of semantic similarity techniques for a corpus of CSCL chats. In *RoEduNet 2015* (pp. 178-183): IEEE.

Holmer, T., Kienle, A., & Wessner, M. (2006). Explicit Referencing in Learning Chats: Needs and Acceptance. In W. Nejdl & K. Tochtermann (Eds.), *EC-TEL 2006* (pp. 170– 184). Crete, Greece: Springer.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). Dirichlet and Inverted Dirichlet Distributions *Continuous Multivariate Distributions* (Vol. 1: Models and Applications, pp. 485–527). New York, NY: Wiley.

Kullback, S., & Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics, 22*(1), 79–86.

Landauer, T. K., & 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.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for wordsense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 265–283). Cambridge, MA: MIT Press.

Lemaire, B. (2009). Limites de la lemmatisation pour l'extraction de significations. In *9es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2009)* (pp. 725–732). Lyon, France: Presses Universitaires de Lyon.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, MA: ACL.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). *Corpus-based and knowledge-based measures of text semantic similarity*. Paper presented at the 21st Int. Conf. AAAI, Boston, Massachusetts.

Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Rebedea, T., Chiru, C.-G., & Gutu, G. (2014). How useful are semantic links for the detection of implicit references in CSCL chats? In *RoEduNet Conference 2014* (pp. 1-6): IEEE.

Rebedea, T., & Gutu, G. (2013). Detecting Implicit References in Chats Using Semantics *Scaling up Learning for Sustained Impact* (pp. 627-628): Springer.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, UK: Cambridge University Press.

Stahl, G. (2006). *Group cognition. Computer support for building collaborative knowledge*. Cambridge, MA: MIT Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association, 101*, 1566–1581.

Trausan-Matu, S. (2010). Computer support for creativity in small groups using chats. *Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information, 3*(2), 81–90.

Trausan-Matu, S., & Rebedea, T. (2010). A polyphonic model and system for inter-animation analysis in chat conversations with multiple participants. In A. F. Gelbukh (Ed.), *11th Int. Conf. Computational Linguistics and Intelligent Text Processing (CICLing 2010)* (pp. 354–363). New York: Springer.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics, ACL '94* (pp. 133–138). New Mexico, USA: ACL.

Acknowledgments

This research was partially supported by the FP7 2008-212578 LTfLL project, by the 644187 EC H2020 *Realising an Applied Gaming Eco-system* (RAGE) project, as well as by University Politehnica of Bucharest through the "Excellence Research Grants" Programs UPB–GEX 12/26.09.2016 and 13/30.09.2016.