Semantic Similarity Applied to Spoken Dialogue Summarization

Iryna Gurevych and **Michael Strube**
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/english/homes/{gurevych|strube}

Abstract

We present a novel approach to spoken dialogue summarization. Our system employs a set of semantic similarity metrics using the noun portion of WordNet as a knowledge source. So far, the noun senses have been disambiguated manually. The algorithm aims to extract utterances carrying the essential content of dialogues. We evaluate the system on 20 Switchboard dialogues. The results show that our system outperforms LEAD, RANDOM and TF*IDF baselines.

1 Introduction

Research in automatic text summarization began in the late 1950s and has been receiving more attention again over the last decade. The maturity of this research area is indicated by recent large-scale evaluation efforts (Radev et al., 2003). In comparison, speech summarization is a rather new research area which emerged only a few years ago. However, the demand for speech summarization is growing because of the increasing availability of (digitally encoded) speech databases (e.g. spoken news, political speeches).

Our research is concerned with the development of a system for automatically generating summaries of conversational speech. As a potential application we envision the automatic generation of meeting minutes. The approach to spoken dialogue summarization presented herein unifies corpus- and knowledge-based approaches to summarization, i.e. we develop a shallow knowledge-based approach. Our system employs a set of semantic similarity metrics which utilize WordNet as a knowledge source. We claim that semantic similarity between a given utterance and the dialogue as a whole is an appropriate criterion for the selection of utterances which carry the essential content of the dialogue, i.e. *relevant* utterances. – In order to study the performance of semantic similarity methods, we remove the noise from the pre-processing modules by manually disambiguating lexical noun senses.

In Section 2, we briefly describe research on summarization and how spoken dialogue summarization differs from text summarization. Section 3 gives the semantic similarity metrics we use and describes how they are applied to the summarization problem. Section 4 provides information about the data used in our experiments, while Section 5 describes the experiments and the results together with their statistical significance.

2 Text, Speech and Dialogue Summarization

Most research on automatic summarization dealt with written text. This work was based either on corpus-based, statistical methods or on knowledge-based techniques (for an overview over both strands of research see Mani & Maybury (1999)). Recent advances in text summarization are mostly due to statistical techniques with some additional usage of linguistic knowledge, e.g. (Marcu, 2000; Teufel & Moens, 2002), which can be applied to unrestricted input.

Research on speech summarization focused mainly on single-speaker, *written-to-be-spoken* text (e.g. spoken news, political speeches, etc.). The methods were mostly derived from work on text summarization, but extended it by exploiting particular characteristics of spoken language, e.g. acoustic confidence scores or intonation. Difficulties arise because speech recognition systems are not perfect. Therefore, spoken dialogue summarization systems have to deal with errors in the input. There are no sentence boundaries in spoken language either.

Work on spoken dialogue summarization is still in its infancy (Reithinger et al., 2000; Zechner, 2002). Multiparty dialogue is much more difficult to process than written text. In addition to the difficulties speech summarization has to face, spoken dialogue contains a whole range of dialogue phenomena as disfluencies, hesitations, interruptions, etc. Also, the information to be summarized may be contributed by different speakers (e.g. in question-answer pairs). Finally, the language used in spoken

dialogue differs from language used in texts. Because discourse participants are able to immediately clarify misunderstandings, the language used does not have to be that explicit.

3 Semantic Similarity

3.1 Semantic Similarity Metrics

Experiments reported here employed Ted Pedersen's (2002) semantic similarity package. We applied five of the metrics, which rely on WordNet as a knowledge base and were developed in the context of work on word sense disambiguation. The first measure is Leacock and Chodorow's (1998) Normalized Path Length (we will refer to it as *lch*). Semantic similarity sim between words w_1 and w_2 is defined as given in Equation 1:

$$sim_{c_1,c_2} = -log\frac{len(c_1,c_2)}{2 \times D} \quad (1)$$

c_1 and c_2 are concepts corresponding to w_1 and w_2.[1] $len(c_1, c_2)$ is the length of the shortest path between them. D is the maximum depth of the taxonomy.

The following measures incorporate an additional, qualitatively different knowledge source based on some kind of corpus analysis. The extended gloss overlaps measure introduced by Banerjee & Pedersen (2003) (referred to as *lesk* in the following) is based on the number of shared words (overlaps) in the WordNet definitions (glosses) of the respective concepts. It also extends the glosses to include the definitions of concepts related to the concept under consideration based on the WordNet hierarchy. Formally, semantic relatedness sim between words w_1 and w_2 is defined by the following equation:

$$sim_{c_1,c_2} = \sum score(R_1(c_1), R_2(c_2)) \quad (2)$$

where R is a set of semantic relations, $score()$ is a function accepting two glosses as input, finding overlaps between them, and returning a corresponding relatedness score.

The remaining three methods require an additional knowledge source, an *information content file* (ICF). This file contains information content values for WordNet concepts, which are needed for computing the semantic similarity score for two concepts. Information content values are based on the frequency counts for respective concepts. Resnik (1995) (*res* for short) calculates the information content of the concept that subsumes the given two concepts in the taxonomy (see Equation 3):

$$sim_{c_1,c_2} = \max_{c \in S(c_1,c_2)} [-\log p(c)] \quad (3)$$

where $S(c_1, c_2)$ is the set of concepts which subsume both c_1 and c_2 and $-\log p(c)$ is the negative log likelihood (information content). The probability p is computed as the relative frequency of the concept. Resnik's measure is based on the intuition that the semantic similarity between concepts may be quantified on the basis of information shared between them. In this case the WordNet hierarchy is used to determine the closest super-ordinate of a pair of concepts.

Jiang & Conrath (1997) proposed to combine edge- and node-based techniques in counting the edges and enhancing it by the node-based calculation of the information content as introduced by Resnik (1995) (the method is abbreviated as *jcn*). The distance between two concepts c_1 and c_2 is formalized as given in Equation 4:

$$dist_{c_1,c_2} = IC_{(c_1)} + IC_{(c_2)} - 2 \times IC(lso(c_1, c_2)) \quad (4)$$

where IC is the information content value of the concept, and $lso(c_1, c_2)$ is the closest subsumer of the two concepts.

The last method is that of Lin (1998) (we call this metric *lin*). He defined semantic similarity using a formula derived from information theory. This measure is sometimes called a universal semantic similarity measure as it is supposed to be application-, domain-, and resource independent. According to this method, the similarity is given in Equation 5:

$$sim_{c_1,c_2} = \frac{2 \times \log p(lso(c_1, c_2))}{\log p(c_1) + \log p(c_2)} \quad (5)$$

3.2 Semantic Similarity in Summarization

The process of automatic dialogue summarization, as defined in the context of this work, means to extract the most relevant utterances from the dialogue. We restate this as a classification problem, which is similar to the definition given by Kupiec et al. (1995). This means that utterances are classified as *relevant* or *irrelevant* for the summary of a specific dialogue. By relevant utterances we mean those carrying the most essential parts of the dialogue's content. The summarization task is, then, to extract the set of utterances from the transcript, which a human would use to make a dialogue summary.

The key idea behind the algorithm presented here is to quantify the degree of semantic similarity between a given utterance and the whole dialogue. We argue that semantic similarity between an utterance

[1] This also refers to the rest of the methods.

A.:	Utt_1	Okay.
	Utt_2	Tell me about your home.
B.:	Utt_3	Well, it's an older home.
	Utt_4	It was made back in the early sixties.
	Utt_5	It's a pier beam house.
A.:	Utt_6	Huh-uh.
B.:	Utt_7	Got three bedrooms, one bath
	Utt_8	and that just makes me scream.
A.:	Utt_9	That's pretty tough.
	Utt_{10}	What area do you live in?
B.:	Utt_{11}	I live in Houston.

Table 1: Switchboard dialogue fragment

Number	Concepts	Sense Number
CR_{Utt_1}	—	—
CR_{Utt_2}	*home*	2
CR_{Utt_3}	*home*	2
CR_{Utt_4}	*sixties*	1
CR_{Utt_5}	*pier, beam, house*	2, 2, 1
CR_{Utt_6}	—	—
CR_{Utt_7}	*bedrooms, bath*	1, 5
CR_{Utt_8}	—	—
CR_{Utt_9}	—	—
$CR_{Utt_{10}}$	*area*	1
$CR_{Utt_{11}}$	*Houston*	1

Table 2: Utterances mapped to WordNet concepts

and the dialogue as a whole represents an appropriate criterion for the selection of relevant utterances. We describe each of the processing steps, employing the example dialogue D from Table 1. This example consists of the set of utterances $\{Utt_1$,..., $Utt_{11}\}$.

3.2.1 Creating conceptual representations

The semantic similarity algorithms introduced in Section 3.1 operate on the noun portion of WordNet. Our approach to dialogue summarization, as previously stated, is to compute semantic similarity for a given pair $\{Utt_n$,$D\}$. In order to do that, we require a WordNet-based conceptual representation of both Utt_n, i.e. CR_{Utt_n}, and D, i.e. CR_D, and compare them using the semantic similarity measures. Therefore, we map the nouns contained in the utterances to their respective WordNet senses and operate on these representations in the subsequent steps. The results of this operation are shown in Table 2. The number in the last column indicates the disambiguated WordNet sense.

The resulting dialogue representation CR_D will be the set of concepts resulting from adding individual utterance representations, i.e. CR_D = {*home, home, sixties, pier, beam, house, bedrooms, bath, area, Houston*}.

3.2.2 Computing average semantic similarity

For each utterance Utt_n, we create a two-dimensional matrix C with the dimensions $(\#CR_D \times \#CR_{Utt_n})$, where # denotes the number of elements in the set. $C = (c_{ij})_{i=1,...,\#CR_D, j=1,...,\#CR_{Utt_n}}$, see Table 3. Then, we compute the semantic similarity $SS_{score}(i,j)$ employing any of the semantic similarity metrics described above for each pair of concepts. The semantic similarity score SS_{final} for CR_{Utt_n} and CR_D is then defined as the average pairwise semantic similarity between all concepts in CR_{Utt_n} and CR_D:

$$SS_{final} = \frac{\sum_{i=1}^{\#CR_{Utt_n}} \sum_{j=1}^{\#CR_D} SS_{score}(i,j)}{\#CR_{Utt_n} \cdot \#CR_D}$$

Computing SS_{final} results in a list of utterances with scores from the respective scoring methods, Table 4. Note that the absolute utterance scores are taken from the real data, i.e. they have been normalized w.r.t. the conceptual representation for the whole dialogue, and not for the dialogue fragment given in Table 1. The rankings were produced for this specific example to make it more illustrative.

3.2.3 Extracting relevant utterances

In order to produce a summary of the dialogue, the utterances first have to be sorted numerically, i.e. ranked on the basis of their scores, see Table 4 for the results of the ranking procedure.[2] Given a compression rate $COMP$ with the range [1,100], the number of utterances classified as *relevant* by an individual scoring method PN_r is a function of the total number of utterances in the dialogue: $PN_r = (COMP/100) \cdot Number_{total}$.[3] Then, given a specific compression rate $COMP$, the top-ranked PN_r utterances will be automatically classified as *relevant*. – Returning to the example in Table 1, we obtain the summaries given in Table 5.

$COMP$	**Selected Utterances**
20%	I live in Houston. Got three bedrooms, one bath.
35%	I live in Houston. Got three bedrooms, one bath. Tell me about your home. Well, it's an older home.

Table 5: Summaries based on Resnik's measure

[2] If two or more utterances get an equal score, they are ranked according to the order of their occurrence.

[3] Note that this number must be rounded to a natural number.

	home	*home*	*sixties*	*pier*	*beam*	*house*	*bedrooms*	*bath*	*area*	*Houston*
bedrooms	3.8021	3.8021	0	2.5158	2.5158	3.8021	9.3157	5.8706	0.8287	0.8287
bath	3.8021	3.8021	0	2.5158	2.5158	3.8021	5.8706	10.7821	0.8287	0.8287

Table 3: Concept matrix C for Utt_7 from Table 1 based on Resnik's measure

Number	**Utterance**	**Resnik's score**	**Rank**
Utt_1	Okay.	—	8
Utt_2	Tell me about your home.	1.4181106409372	3
Utt_3	Well, it's an older home.	1.4181106409372	4
Utt_4	It was made back in the early sixties.	0.551830914995721	7
Utt_5	It's a pier beam house.	1.18821772523631	6
Utt_6	Huh-uh.	—	9
Utt_7	Got three bedrooms, one bath	1.50689651387565	2
Utt_8	and that just makes me scream.	—	10
Utt_9	That's pretty tough.	—	11
Utt_{10}	What area do you live in?	1.25186984433606	5
Utt_{11}	I live in Houston.	1.51301080520959	1

Table 4: Utterance scores based on Resnik's measure

	Tokens	**Utterances**	**Turns**
Total	34830	3275	1852
Average	1741.5	163.75	92.6

Table 6: Descriptive corpus statistics

4 Data

The data used in the experiments are 20 randomly chosen Switchboard dialogues (Greenberg, 1996). These dialogues contain two-sided telephone conversations among American speakers of at least 10 minutes duration. The callers were given a certain topic for discussion. The recordings of spontaneous speech were, then, transcribed. Statistical data about the corpus, i.e. total numbers and averages for separate dialogues, are given in Table 6. Tokens are defined as running words and punctuation. An utterance is a complete unit of speech spoken by a single speaker, while a turn is a joint sequence of utterances produced by one speaker.

In the annotation experiments, we tested whether humans could reliably determine the utterances conveying the overall meaning of the dialogue. Therefore, each utterance is assumed to be a markable, i.e. the expression to be annotated resulting in a total of 3275 markables in the corpus. Three annotators were instructed to select the most important utterances. They were supposed to first read the dialogue und then to mark about 10% of all utterances in the dialogue as being relevant. Then, we produced two kinds of *Gold Standards* from these data. *Gold Standard 1* included the utterances which were marked by all three annotators as being relevant. *Gold Standard 2* included the utterances which were selected by at least two annotators.

Table 7 shows the results of these experiments. We present the absolute number of markables selected as relevant by separate annotators and in two *Gold Standards*. Also, we indicate the percentage, given the total number of markables 3275. As the table shows, *Gold Standard 1* includes only 3.69% of all markables. Therefore, we used *Gold Standard 2* in the evaluation reported in Section 5. The Kappa coefficient for inter-annotator agreement varied from 0.1808 to 0.6057 for individual dialogues. An examination of the particular dialogue with the very low Kappa rate showed that this was one of the shortest ones. It did not have a well-defined topical structure, resulting in a low agreement rate between annotators. For the whole corpus, the Kappa coefficient yielded 0.4309. While this is not a high agreement rate on a general scale, it is comparable to what has been reported concerning the task of summarization and in particular dialogue summarization.

5 Evaluation

5.1 Evaluation Metrics and Baselines

We reformulated the problem in terms of standard information retrieval evaluation metrics: $Precision = PP/PNr$, $Recall = PP/NP$, and $Fmeasure = 2 \cdot Prec \cdot Rec/(Prec + Rec)$. PP is the number of cases where the individual scoring method and the *Gold Standard* agree. PNr is computed according to the definition given in Section 3.

	Annotator 1		Annotator 2		Annotator 3		Gold Standard 1		Gold Standard 2	
Σ	417	**12.73%**	350	**10.69%**	347	**10.6%**	121	**3.69%**	310	**9.47%**

Table 7: Number of markables labeled as *relevant*

NP is the total number of utterances marked as *relevant* in the *Gold Standard*. For comparison, three baseline systems were implemented. The first system is the RANDOM baseline, where relevant utterances (depending on the compression rate) were selected by chance. The second baseline system is based on the TF*IDF scoring metric. A large corpus is required to make this method fully powerful. Therefore, we computed TF*IDF scores for every word on the basis of 2431 Switchboard dialogues (ca. 19.3 MB of ASCII text). Then, an average TF*IDF score for each utterance of the 20 dialogues in our corpus was computed by adding the individual scores for all words in the utterance and normalizing by the number of words. The LEAD baseline is based on the intuition that the most important utterances tend to occur at the beginning of the discourse. While this observation is true for the domain of news, the LEAD baseline is not necessarily efficient for the genre of spontaneous dialogues. However, given the Switchboard experimental data collection setup, the dialogues usually directly start with the discussions of the topic. This hypothesis was supported by evidence from our own annotation experiments, too.

5.2 Results

Experiments were performed using the semantic similarity package V0.05 (Pedersen, 2002) and WordNet 1.7.1. We employed *Gold Standard 2* (see Section 4). Three of the methods, namely *res*, *lin*, *jcn*, require the *information content file* (ICF). A method for computing the information content of concepts from large corpora of text is given in Resnik (1995). ICF contains a list of synsets along with their part of speech and frequency count. We compare the results obtained with 2 different ICFs:

- a WordNet-based ICF, provided at the time of the installation of the similarity package with pre-computed frequency values on the basis of WordNet (WD_ICF);
- an ICF, generated specifically on the basis of 2431 Switchboard dialogues with the help of utilities distributed together with the similarity package (SW_ICF).

Figures 1 and 2 indicate the performance of all methods in terms of F-measure. The results of the

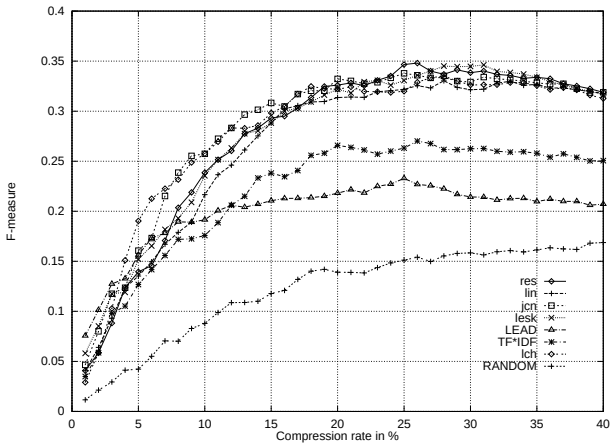

Figure 1: Results based on WordNet ICF

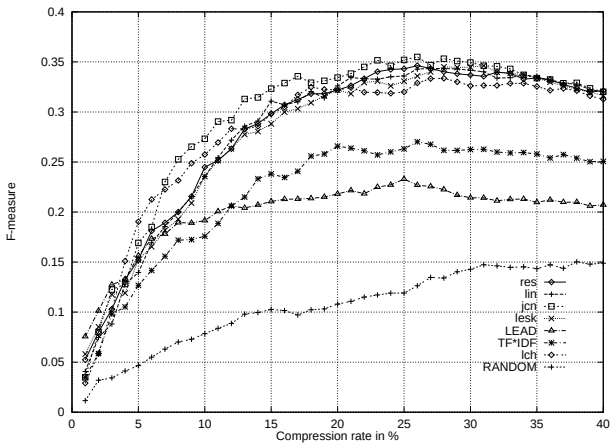

Figure 2: Results based on Switchboard ICF

semantic similarity methods making use of the information content file generally improve when the Switchboard-based ICF is used. The improvements are especially significant for the *jcn* and *lin* measures, while this does not seem to be the case for the *res* measure (depending on a specific compression rate).

The summarization methods perform best for the compression rates in the interval [20,30]. Given these rates and the Switchboard-based ICF, the competing methods display the following performance (in descending order): *jcn*, *res*, *lin*, *lesk*, *lch*, *tf*idf*, *lead*, *random*. For the default ICF the picture is

slightly different: *res*, *jcn* and *lesk*, *lch*, *lin*, *tf*idf*, *lead*, *random* (see Table 8). *lch* relying on WordNet structure only performs worse than the rest of similarity metrics incorporating some corpus evidence.

A direct comparison of our evaluation with alternative results, e.g., Zechner's (2002) is problematic. Though Zechner's results are based on Switchboard, too, he employs a different evaluation scheme. The evaluation is broken down to the word level. The results are compared with multiple human annotations instead of a *Gold Standard*.

5.3 Statistical Significance and Error Analysis

For determining whether there is a significant difference between the summarization approaches pairwise, we use a paired related *t-test* (as the parent distribution is unknown). The null hypothesis states there is no difference between the two distributions. On consulting the *t-test* tables, we obtain the significance values presented in Table 9, given the compression rate 25%[4] and the Switchboard ICF. These results indicate that there is no statistically significant difference in the performance between the *res*, *lin*, *jcn* and *lesk* methods. However, all of them significantly outperform the LEAD, TF*IDF and RANDOM baselines.

The maximum Recall of the semantic similarity-based summarization methods in the current implementation is limited to about 90%, given $COMP = 100\%$. This means that if the system compiled a 100% "summary", it would miss 10% of all utterances marked as *relevant*. The reason lies in the fact that the algorithm operates on the concepts created by mapping nouns to their WordNet senses. Thus, the *relevant* utterances which do not have nouns on the surface, but contain for example anaphorical expressions realized as pronouns, are missed in the input. Resolving anaphorical expressions in the preprocessing stage may eliminate this error source.

6 Concluding Remarks

We introduced a new approach to spoken dialogue summarization. Our approach combines statistical, i.e. corpus-based, and knowledge-based techniques. It utilizes the knowledge encoded in the noun part of WordNet and applies a set of semantic similarity metrics to dialogue summarization. All semantic similarity-based summarization methods outperform RANDOM, LEAD and TF*IDF baseline systems. In the following, we discuss some remaining challenges and future research.

More sophisticated data pre-processing. We plan to incorporate the pre-processing components used by Zechner (2002) and evaluate their contribution to our task. Including an anaphora resolution component would also result in better Recall.

Automatic word sense disambiguation. Switchboard conversational speech is highly ambiguous. Automatic disambiguation of noun senses to WordNet concepts is important in order to integrate our approach into real-life summarization systems.

Investigating other types of information in parallel. A clear desideratum will be assessing the overall coherence of the discourse, speaker info, turn type, information about non-nouns.

Application to text and speech summarization. Our approach can be applied to *written-to-be-spoken* speech and text summarization. It will be interesting to investigate whether conceptual structures of texts (the input to our system) are comparable to the conceptual structures found in dialogues.

Readability, coherence, and usability of the summaries produced. A close examination of summaries based on human comprehension will be interesting. It may be necessary to introduce filtering or other post-processing techniques improving the quality of summaries.

Even without very sophisticated pre-processing of the dialogue data, our algorithm yields promising results. It was evaluated on the Switchboard data, which is a challenging evaluation corpus. Our vision is to adopt the summarization approach presented here in a system used for the automatic production of meeting minutes.

[4]Roughly speaking, the differences are most evident for compression rates between 20% and 30%.

Acknowledgments

This work has been funded by the Klaus Tschira Foundation. We thank Christoph Zwirello for his valuable contributions, the annotators Tatjana Medvedeva, Vanessa Michelli and Iryna Zhmaka, and Ted Pederson and colleagues for their software.

References

Banerjee, S. & T. Pedersen (2003). Extended gloss overlap as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 9–15 August, 2003, pp. 805–810.

Greenberg, S. (1996). The Switchboard transcription project. In *Proceedings of the Large Vocabulary Continuous Speech Recognition Summer Research Workshop*, Baltimore, Maryland, USA, April 1996.

Jiang, J. J. & D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING)*, pp. 19–33. Tapei, Taiwan.

		RANDOM	LEAD	TF*IDF	Res	Lin	Jcn	Lesk	Lch
Precision 10%	WD_ICF	.08563	.18654	.17125	.23242	.21101	.25076	.22936	.25076
	SW_ICF	.07645			.23853	.22936	.26606		
20%	WD_ICF	.1026	.16159	.19512	.23933	.23018	.24390	.23780	.23780
	SW_ICF	.07963			.23780	.24085	.24695		
30%	WD_ICF	.10429	.14140	.17192	.22380	.21261	.21668	.22584	.21567
	SW_ICF	.09407			.22075	.22482	.23093		
Recall 10%	WD_ICF	.09032	.19677	.18065	.24516	.22258	.26452	.24194	.26452
	SW_ICF	.08065			.25161	.24194	.28065		
20%	WD_ICF	.21613	.34194	.41290	.50645	.48710	.51613	.50323	.50323
	SW_ICF	.16774			.50323	.50968	.52258		
30%	WD_ICF	.32903	.44839	.54516	.70968	.67419	.68710	.71613	.68387
	SW_ICF	.29677			.70000	.71290	.73226		
F-measure 10%	WD_ICF	.08791	.19152	.17582	.23862	.21664	.25746	.23548	.25746
	SW_ICF	.07849			.24490	.23548	.27316		
20%	WD_ICF	.13915	.21946	.26501	.32505	.31263	.33126	.32298	.32298
	SW_ICF	.108			.32298	.32712	.33540		
30%	WD_ICF	.15839	.21500	.26141	.34029	.32328	.32947	.34339	.32792
	SW_ICF	.14286			.33565	.34184	.35112		

Table 8: Precision, Recall and F-measure for 10%, 20% and 30% and two ICFs

	res	*lin*	*jcn*	*lesk*	*lead*	*tf*idf*	*lch*	*random*
res	XXX	p>0.05	p>0.05	p>0.05	**p<0.01**	**p<0.01**	p>0.05	**p<0.01**
lin		XXX	p>0.05	p>0.05	**p<0.01**	**p<0.05**	p>0.05	**p<0.01**
jcn			XXX	p>0.05	**p<0.01**	**p<0.01**	**p<0.05**	**p<0.01**
lesk				XXX	**p<0.01**	**p<0.05**	p>0.05	**p<0.01**
lead					XXX	p>0.05	**p<0.01**	**p<0.01**
*tf*idf*						XXX	p>0.05	**p<0.01**
lch							XXX	**p<0.01**
random								XXX

Table 9: Statistical significance of results at $COMP$=25% and based on SW_ICF

Kupiec, J., J. O. Pedersen & F. Chen (1995). A trainable document summarizer. In *Research and Development in Information Retrieval*, pp. 68–73.

Leacock, C. & M. Chodorow (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, pp. 265–283. Cambridge: MIT Press.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304. Morgan Kaufmann, San Francisco, CA.

Mani, I. & M. T. Maybury (Eds.) (1999). *Advances in Automatic Text Summarization*. Cambridge/MA, London/England: MIT Press.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge/MA: The MIT Press.

Pedersen, T. (2002). *Semantic Similarity Package*. http://www.d.umn.edu/∼tpederse/similarity.html.

Radev, D. R., S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. Celebi, D. Liu & E. Drabek (2003). Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 7–12 July 2003, pp. 375–382.

Reithinger, N., M. Kipp, R. Engel & J. Alexandersson (2000). Summarizing multilingual spoken negotiation dialogues. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, 1–8 August 2000, pp. 310–317.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montréal, Canada, 1995, Vol. 1, pp. 448–453.

Teufel, S. & M. Moens (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Zechner, K. (2002). Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.