		Rater 2	
		Positive	Negative
Rater 1	Positive	a	b
	Negative	c	d

Table 1. Possible choices for any two raters

that sentence which contains the first appearance of a chain member. Doran et al. (2004) sum up the weights all words in the sentence, which correspond to the chain weights in which these words occur. We choose the latter heuristic because it significantly outperforms the former method in our experiments.

The highest scoring sentences from the document, presented in their original order, form the automatically generated summary. How many sentences are extracted depends on the requested summary length, which is defined as the percentage of the document length.

4.2 Experimental Settings

For evaluation we used a subset of a manually annotated corpus specifically created to evaluate text summarization systems (Hasler et al. 2003). We concentrate only on documents with at least two manually produced summaries: 11 science and 29 newswire articles with two summaries each, and 7 articles additionally annotated by a third person. This data allows us to compare the consistency of the system with humans to their consistency with each other.

The results are evaluated with the Kappa statistic κ, defined for Table 1 as follows:

$$\kappa = \frac{2(ab - bc)}{(a+c)(c+9)+(b+d)(a+b)} \quad (3)$$

It takes into account the probability of chance agreement and is widely used to measure inter-rater agreement (Hripcsak and Rothshild, 2005). The ideal automatic summarization algorithm should have as high agreement with human subjects as they have with each other.

We also use a baseline approach (BL) to estimate the advantage of using the proposed lexical chaining algorithm (LCA). It extracts text summaries in exactly the manner described in Section 4.1, with the exception of the lexical chaining stage. Thus, when weighting sentences, the frequencies of *all* WordNet mappings are taken into account without the implicit word sense disambiguation provided by lexical chains.

		Humans	BL	LCA
29 newswire articles	S1	0.32	0.19	0.20
	S2		0.20	0.24
11 science articles	S1	0.34	0.08	0.13
	S2		0.13	0.22

Table 2. Kappa agreement on 40 summaries

	vs. human 2,3 and 1	vs. BL	vs. LCA
human 1	0,41	0,30	0,30
human 2	0,38	0,22	0,24
human 3	0,28	0,17	0,24
average	0,36	0,23	0,26

Table 3. Kappa agreement on 7 newswire articles

4.3 Results

Table 2 compares the agreement among the human annotators and their agreement with the baseline approach BL and the lexical chain algorithm LCA. The agreement between humans is low, which confirms that sentence extraction is a highly subjective task. The lexical chain approach LCA significantly outperforms the baseline BL, particularly on the science articles.

While the average agreement of the LCA with humans is still low, the picture changes when we look at the agreement on individual documents. Human agreement varies a lot (*stdev* = 0.24), while results produced by LCA are more consistent (*stdev* = 0.18). In fact, for over 50% of documents LCA has greater or the same agreement with one or both human annotators than they with each other. The overall superior performance of humans is due to exceptionally high agreement on a few documents, whereas on another couple of documents LCA failed to produce a consistent summary with both subjects. This finding is similar to the one mentioned by Silber and McCoy (2002).

Table 3 shows the agreement values for 7 newswire articles that were summarized by three human annotators. Again, LCA clearly outperforms the baseline BL. Interestingly, both systems have a greater agreement with the first subject than the first and the third human subjects with each other.

5 Lexical Chains for Keyphrase Indexing

Keyphrase indexing is the task of identifying the main topics in a document. The drawback of conventional indexing systems is that they analyze