Towards robust context-sensitive sentence alignment for monolingual corpora

Towards Robust Context-Sensitive Sentence Alignment for Monolingual Corpora

Rani Nelken and **Stuart M. Shieber**
Division of Engineering and Applied Sciences
Harvard University
33 Oxford St.
Cambridge, MA 02138
{nelken,shieber}@deas.harvard.edu

Rani Nelken and Stuart M. Shieber. 2006. Towards Robust Context-Sensitive Sentence Alignment for Monolingual Corpora. In *Proceedings of the 11th Conference of the Euorpean Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy, 3-7 April.

Abstract

Aligning sentences belonging to comparable monolingual corpora has been suggested as a first step towards training text rewriting algorithms, for tasks such as summarization or paraphrasing. We present here a new monolingual sentence alignment algorithm, combining a sentence-based TF*IDF score, turned into a probability distribution using logistic regression, with a global alignment dynamic programming algorithm. Our approach provides a simpler and more robust solution achieving a substantial improvement in accuracy over existing systems.

1 Introduction

Sentence-aligned bilingual corpora are a crucial resource for training statistical machine translation systems. Several authors have suggested that large-scale aligned monolingual corpora could be similarly used to advance the performance of monolingual text-to-text rewriting systems, for tasks including summarization (Knight and Marcu, 2000; Jing, 2002) and paraphrasing (Barzilay and Elhadad, 2003; Quirk et al., 2004). Unlike bilingual corpora, such as the Canadian Hansard corpus, which are relatively rare, it is now fairly easy to amass corpora of related monolingual documents. For instance, with the advent of news aggregator services such as "Google News", one can readily collect multiple news stories covering the same news item (Dolan et al., 2004). Utilizing such a resource requires aligning related documents at a finer level of resolution, identifying which sentences from one document align with which sentences from the other.

Previous work has shown that aligning related monolingual documents is quite different from the well-studied multi-lingual alignment task. Whereas documents in a bilingual corpus are typically very closely aligned, monolingual corpora exhibit a much looser level of alignment, with similar content expressed using widely divergent wording, grammatical form, and sentence order. Consequently, many of the simple surface-based methods that have proven to be so successful in bilingual sentence alignment, such as correlation of sentence length, linearity of alignment, and a predominance of one-to-one sentence mapping, are much less likely to be effective for monolingual sentence alignment.

Barzilay and Elhadad (2003) suggested that these disadvantages could be at least partially offset by the recurrence of the same lexical items in document pairs. Indeed, they showed that a simple cosine word-overlap score is a good baseline for the task, outperforming much more sophisticated methods. They also observed that context is a powerful factor in determining alignment. They illustrated this on a corpus of Encyclopedia Britannica entries describing world cities, where each entry comes in two flavors, the comprehensive encyclopedia entry, and a shorter and simpler elementary version. Barzilay and Elhadad used context in two different forms. First, using inter-document context, they took advantage of commonalities in the topical structure of the encyclopedia entries to identify paragraphs that are likely to be about the same topic. They then took advantage of intra-document context by using dynamic programming to locally align sequences of sentences belonging to paragraphs about the same topic, yielding improved accuracy on the corpus. While powerful, such commonalities in document structure appear to be a special feature of the Britannica corpus, and therefore cannot be relied upon for other corpora.

In this paper we present a novel algorithm for sentence alignment in monolingual corpora. At the core of the algorithm is a classical similar-

ity score based on differentially weighting words according to their Term Frequency-Inverse Document Frequency (TF*IDF) (Spärck-Jones, 1972; Salton and Buckley, 1988). We treat sentences as documents, and the collection of sentences in the two documents being compared as the document collection, and use this score to estimate the probability that two sentences are aligned using logistic regression. Surprisingly, this approach by itself yields competitive accuracy, yielding the same level of accuracy as Barzilay and Elhadad's algorithm, and higher than all previous approaches on the Britannica corpus. Such matching, however, is still noisy. We further improve accuracy by using a global alignment dynamic programming algorithm, which prunes many spurious matches.

Our approach validates Barzilay and Elhadad's observation regarding the utility of incorporating context. In fact, we are able to extract more information out of the intra-document context. First, by using TF*IDF at the level of sentences, we weigh words in a sentence with respect to other sentences of the document. Second, global alignment takes advantage of (noisy) linear order of sentences. We make no use of inter-document context, and in particular make no assumptions about common topical structure that are unique to the Britannica corpus, thus ensuring the scalability of the approach.

Indeed, we successfully apply our algorithm to a very different corpus, the three *Synoptic gospels* of the New Testament: Matthew, Mark, and Luke. Putting aside any religious or theological significance of these texts, they offer an excellent data source for studying alignment, since they contain many parallels, which have been conveniently annotated by bible scholars (Aland, 1985). Our algorithm achieves a significant improvement over the baseline for this corpus as well, demonstrating the general applicability of our approach.

2 Related work

Several authors have tackled the monolingual sentence correspondence problem. SimFinder (Hatzivassiloglou et al., 1999; Hatzivassiloglou et al., 2001) examined 43 different features that could potentially help determine the similarity of two short text units (sentences or paragraphs). Of these, they automatically selected 11 features, including word overlap, synonymy as determined by WordNet (Fellbaum, 1998), matching proper nouns and noun phrases, and sharing semantic classes of verbs (Levin, 1993).

The Decomposition method (Jing, 2002) relies on the observation that document summaries are often constructed by extracting sentence fragments from the document. It attempts to identify such extracts, using a Hidden Markov Model of the process of extracting words. The HMM uses features of word identity and document position, in which transition probabilities are based on locality assumptions. For instance, after a word is extracted, an adjacent word or one that belongs to a nearby sentence is more likely to be extracted than one that is further away.

Barzilay and Elhadad (2003) apply a 4-step algorithm:

1. Cluster the paragraphs of the training documents into topic-specific clusters, based on word overlap. For instance, paragraphs in the Britannica city entries describing climate might cluster together.

2. Learn mapping rules between paragraphs of the full and elementary versions, taking the word-overlap and the clusters as features.

3. Given a new pair of texts, identify sentence pairs with high overlap, and take these to be aligned. Then, classify paragraphs according to the clusters learned in Step 1, and use the mapping rules of Step 2 to match pairs of paragraphs between the documents.

4. Finally, take advantage of the paragraph clustering and mapping, by locally aligning only sentences belonging to mapped paragraph pairs.

Dolan et al. (2004) used Web-aggregated news stories to learn both sentence-level and word-level alignments. Having collected a large corpus of clusters of related news stories from Google and MSN news aggregator services, they first seek related sentences, using two methods. First, using a high Levenshtein distance score they identify 139K sentence pairs of which about 16.7% are estimated to be unrelated (using human evaluation of a sample). Second, assuming that the first two sentences of related news stories should be matched, provided they have a high enough word-overlap, yields 214K sentence pairs of which about 40% are estimated to be unrelated. No recall estimates

are provided; however, with the release of the annotated Microsoft Research Paraphrase Corpus,[1] it is apparent that Dolan et al. are seeking much more tightly related pairs of sentences than Barzilay and Elhadad, ones that are virtually semantically equivalent. In subsequent work, the same authors (Quirk et al., 2004) used such matched sentence pairs to train Giza++ (Och and Ney, 2003) on word-level alignment.

The recent PASCAL "Recognizing Textual Entailment" (RTE) challenge (Dagan et al., 2005) focused on the problem of determining whether one sentence entails another. Beyond the difference in the definition of the required relation between sentences, the RTE challenge focuses on isolated sentence pairs, as opposed to sentences within a document context. The task was judged to be quite difficult, with many of the systems achieving relatively low accuracy.

3 Data

The Britannica corpus, collected and annotated by Barzilay and Elhadad (2003), consists of 103 pairs of comprehensive and elementary encyclopedia entries describing major world cities. Twenty of these document pairs were annotated by human judges, who were asked to mark sentence pairs that contain at least one clause expressing the same information, and further split into a training and testing set.

As a rough indication of the diversity of the dataset and the difference of the task from bilingual alignment, we define the *alignment diversity measure* (*ADM*) for two texts, T_1, T_2, to be: $\frac{2*matches(T_1,T_2)}{|T_1|+|T_2|}$, where *matches* is the number of matching sentence pairs. Intuitively, for closely aligned document pairs, as prevalent in bilingual alignment, one would expect an *ADM* value close to 1. The average *ADM* value for the training document pairs of the Britannica corpus is 0.26.

For the gospels, we use the King James version, available electronically from the Sacred Text Archive.[2] The gospels' lengths span from 678 verses (Mark) to 1151 verses (Luke), where we treat verses as sentences. For training and evaluation purposes, we use the list of parallels given by Aland (1985).[3] We use the pair Matthew-Mark for training and the two pairs: Matthew-Luke and Mark-Luke for testing. Whereas for the Britannica corpus parallels were marked at the resolution of sentences, Aland's annotation presents parallels as matched sequences of verses, known as *pericopes*. For instance, Matthew:4.1-11 matches Mark:1.12-13. We write $v \in p$ to indicate that verse v belongs to pericope p.[4]

4 Algorithm

We now describe the algorithm, starting with the TF*IDF similarity score, followed by our use of logistic regression, and the global alignment.

4.1 From word overlap to TF*IDF

Barzilay and Elhadad (2003) use a cosine measure of word-overlap as a baseline for the task. As can be expected, word overlap is a relatively effective indicator of sentence similarity and relatedness (Marcu, 1999). Unfortunately, plain word-overlap assigns all words equal importance, not even distinguishing between function and content words. Thus, once the overlap threshold is decreased to improve recall, precision degrades rapidly. For instance, if a pair of sentences has one or two words in common, this is inconclusive evidence of their similarity or difference.

One way to address this problem is to differentially weight words using the TF*IDF scoring scheme, which has become standard in Information Retrieval (Salton and Buckley, 1988). IDF was also used for the similar task of directional entailment by Monz and de Rijke (2001). To apply this scheme for the task at hand we diverge from the standard IDF definition by viewing each sentence as a document, and the pair of documents as a combined collection of N single-sentence documents. For a term t in sentence s, we define $\mathrm{TF_s(t)}$ to be a binary indicator of whether t occurs in s,[5] and $\mathrm{DF(t)}$ to be the number of sentences in which t occurs. The TF*IDF weight is:

$$w_s(t) =_{def} \mathrm{TF}_s(t) \cdot \log\left(\frac{N}{\mathrm{DF}(t)}\right) \quad .$$

[1] http://research.microsoft.com/research/downloads/

[2] http://www.sacred-texts.com

[3] The parallels are available online from http://www.bible-researcher.com/parallels.html.

[4] The annotation of matched pericopes induces a partial segmentation of each gospel into paragraph-like segments. Since this segmentation is part of the gold annotation, we do not use it in our algorithm.

[5] Using a binary indicator rather than the more typical number of occurrences yielded better accuracy on the Britannica training set. This is probably due to the "documents" being only of sentence length.

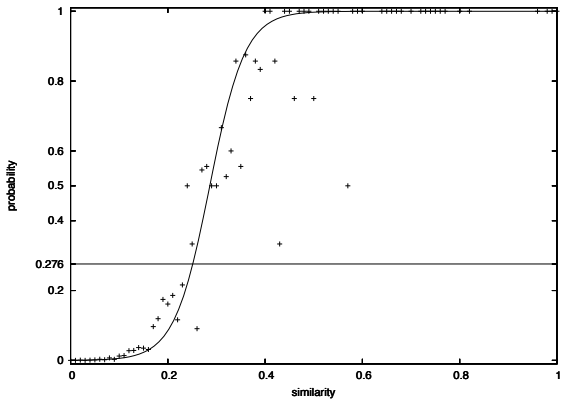

Figure 1: Logistic Regression for Britannica training data

We use these scores as the basis of a standard cosine similarity measure,

$$sim(s_1, s_2) = \frac{s_1 \cdot s_2}{|s_1||s_2|} = \frac{\sum_t w_{s_1}(t) \cdot w_{s_2}(t)}{\sqrt{\sum_t w_{s_1}^2(t) \sum_t w_{s_2}^2(t)}} \quad .$$

We normalize terms by using Porter stemming (Porter, 1980). For the Britannica corpus, we also normalized British/American spelling differences using a small manually-constructed lexicon.

4.2 Logistic regression

TF*IDF scores provide a numeric measure of sentence similarity. To use them for choosing sentence pairs, we proceeded to learn a probability of two sentences being matched, given their TF*IDF similarity score, $pr(match = 1 \mid sim)$. We expect this probability to follow a sigmoid-shaped curve. While it is always monotonically increasing, the rate of ascent changes; for very low or very high values it is not as steep as for middle values. This reflects the intuition that while we always prefer a higher scoring pair over a lower scoring pair, this preference is more pronounced in the middle range than in the extremities.

Indeed, Figure 1 shows a graph of this distribution on the training part of the Britannica corpus, where point (x, y) represents the fraction y of correctly matched sentences of similarity x. Overlayed on top of the points is a logistic regression model of this distribution, defined as the function

$$p = \frac{e^{a+bx}}{1 + e^{a+bx}} \quad ,$$

where a and b are parameters. We used Weka (Witten and Frank, 1999) to automatically learn the parameters of the distribution on the training data. These are set to $a = -7.89$ and $b = 27.56$ for the Britannica corpus.

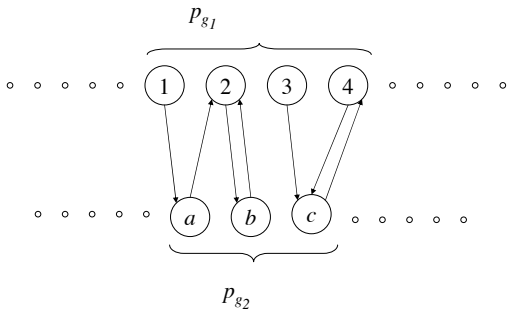

Figure 2: Reciprocal best hit example. Arrows indicate the best hit for each verse. The pairs considered correct are $\langle 2, b \rangle$ and $\langle 4, c \rangle$.

Logistic regression scales the similarity scores monotonically but non-linearly. In particular, it changes the density of points at different score levels. In addition, we can use this distribution to choose a threshold, th, for when a similarity score is indicative of a match. Optimizing the F-measure on the training data using Weka, we choose a threshold value of $th = 0.276$. Note that since the logistic regression transformation is monotonic, the existence of a threshold on probabilities implies the existence of a threshold on the original *sim* scores. Moreover, such a threshold might be obtained by means other than logistic regression. The scaling, however, will become crucial once we do additional calculations with these probabilities in Section 4.4.

Applying logistic regression to the gospels is complicated by the fact that we only have a correct alignment at the resolution of pericopes, and not individual verses. Verse pairs that do not belong to a matched pericope pair can be safely considered unaligned, but for a matched pericope pair, p_{g_1}, p_{g_2}, we do not know which verse is matched with which. We solve this by searching for the *reciprocal best hit*, a method often used to find orthologous genes in related species (Mushegian and Koonin, 1996). For each verse in each pericope, we find the top matching verse in the other pericope. We take as correct all and only pairs of verses x, y, such that x is y's best match and y is x's best match. An example is shown in Figure 2. Taking these pairs as matched yields an *ADM* value of 0.34 for the training pair of documents.

We used the reciprocally best-matched pairs of the training portion of the gospels to find logistic regression parameters ($a = -9.60, b = 25.00$), and

a threshold, ($th = 0.250$). Note that we rely on this matching only for training, but not for evaluation (see Section 5.2).

4.3 Method 1: TF*IDF

As a simple method for choosing sentence pairs, we just select all sentence pairs with $pr(match) > th$. We use the following additional heuristics:

- We unconditionally match the first sentence of one document with the first sentence of the other document. As noted by Quirk et al. (2004), these are very likely to be matched, as verified on our training set as well.
- We allow many-to-one matching of sentences, but limit them to at most 2-to-1 sentences in both directions (by allowing only the top two matches per sentence to be chosen), since such multiple matchings often arise due to splitting a sentence into two, or conversely, merging two sentences into one.

4.4 Method 2: TF*IDF + Global alignment

Matching sentence pairs according to TF*IDF ignores sentence ordering completely. For bilingual texts, Gale and Church (1991) demonstrated the extraordinary effectiveness of a global alignment dynamic programming algorithm, where the basic similarity score was based on the difference in sentence lengths, measured in characters. Such methods fail to work in the monolingual case. Gale and Church's algorithm (using the implementation of Danielsson and Ridings (1997)) yields 2% precision at 2.85% recall on the Britannica corpus. Moore's algorithm (2002), which augments sentence length alignment with IBM Model 1 alignment, reports zero matching sentence pairs (regardless of threshold).

Nevertheless, we expect sentence ordering can provide important clues for monolingual alignment, bearing in mind two main differences from the bilingual case. First, as can be expected by the *ADM* value, there are many gaps in the alignment. Second, there can be large segments that diverge from the linear order predicted by a global alignment, as illustrated by the oval in Figure 3 (Figure 2, (Barzilay and Elhadad, 2003)).

To model these features of the data, we use a variant of Needleman-Wunsch alignment (1970). We compute the optimal alignment between sentences $1..i$ of the comprehensive text and sentences $1..j$ of the elementary version by

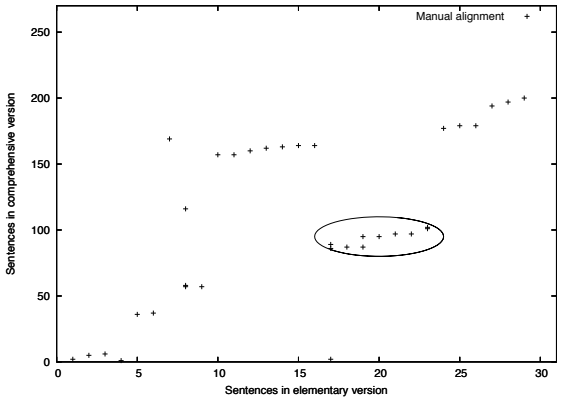

Figure 3: Gold alignment for a text from the Britannica corpus.

$$s(i,j) = \max \begin{cases} s(i-1,j-1) + pr(match(i,j)) \\ s(i-1,j) + pr(match(i,j)) \\ s(i,j-1) + pr(match(i,j)) \end{cases}$$

Note that the dynamic programming sums match probabilities, rather than the original *sim* scores, making crucial use of the calibration induced by the logistic regression. Starting from the first pair of sentences, we find the best path through the matrix indexed by i and j, using dynamic programming. Unlike the standard algorithm, we assign no penalty to off-diagonal matches, allowing many-to-one matches as illustrated schematically in Figure 4. This is because for the loose alignment exhibited by the data, being off-diagonal is not indicative of a bad match. Instead, we prune the complete path generated by the dynamic programming using two methods. First, as in Section 4.3, we limit many-to-one matches to 2-to-1, by allowing just the two best matches per sentence to be included. Second, we eliminate sentence pairs with very low match probabilities ($pr(match) < 0.005$), a value learned on the training data. Finally, to deal with the divergences from the linear order, we add the top n pairs with very high match probability, above a higher threshold, th'. Optimizing on the training data, we set $n = 5$ and $th' = 0.65$ for both corpora.

Note that although Barzilay and Elhadad also used an alignment algorithm, they restricted it only to sentences judged to belong to topically related paragraphs. As noted above, this restriction relies on a special feature of the corpus, the fact that encyclopedia entries follow a relatively regular structure of paragraphs. By not relying on such

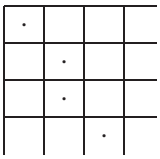

Figure 4: Global alignment

corpus-specific features, our approach gains in robustness.

5 Evaluation

5.1 Britannica corpus

Precision/recall curves for both methods, aggregated over all the documents of the testing portion of the Britannica corpus are given in Figure 5. To obtain different precision/recall points, we vary the threshold above which a sentence pair is deemed matched. Of course, when practically applying the algorithm, we have to pick a particular threshold, as we have done by choosing *th*. Precision/recall values at this threshold are also indicated in the figure.[6]

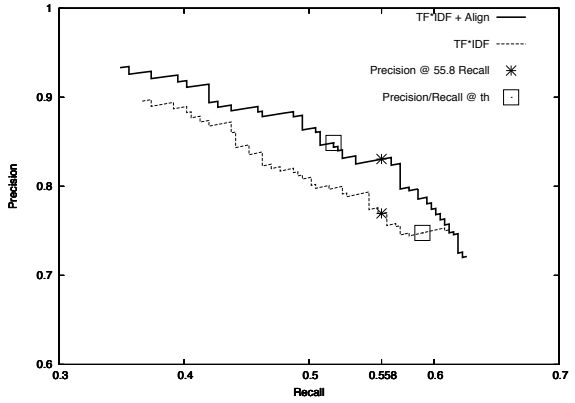

Figure 5: Precision/Recall curves for the Britannica corpus

Comparative results with previous algorithms are given in Table 1, in which the results for Barzilay and Elhadad's algorithm and previous ones are taken from Barzilay and Elhadad (2003). The paper reports the precision at 55.8% recall, since the Decomposition method (Jing, 2002) only produced results at this level of recall, as some of the method's parameters were hard-coded.

Interestingly, the TF*IDF method is highly competitive in determining sentence similarity.

[6]Decreasing the threshold to 0.0 does not yield all pairs, since we only consider pairs with similarity strictly greater than 0.0, and restrict many-to-one matches to 2-to-1.

Algorithm	Precision
SimFinder	24%
Word Overlap	57.9%
Decomposition	64.3%
Barzilay & Elhadad	76.9%
TF*IDF	77.0%
TF*IDF + Align	**83.1%**

Table 1: Precision at 55.8% Recall

Despite its simplicity, it achieves the same performance as Barzilay and Elhadad's algorithm,[7] and is better than all previous ones. Significant improvement is achieved by adding the global alignment.

Clearly, the method is inherently limited in that it can only match sentences with some lexical overlap. For instance, the following sentence pair that should have been matched was missed:

- Population soared, reaching 756,000 by 1903, and urban services underwent extensive modification.

- At the beginning of the 20th century, Warsaw had about 700,000 residents.

Matching "1903" with "the beginning of the 20th century" goes beyond the scope of any method relying predominantly on word identity. The hope is, however, that such mappings could be learned by amassing a large corpus of accurately sentence-aligned documents, and then applying a word-alignment algorithm, as proposed by Quirk et al. (2004). Incidentally, examining sentence pairs with high TF*IDF similarity scores, there are some striking cases that appear to have been missed by the human judges. Of course, we faithfully and conservatively relied on the human annotation in the evaluation, ignoring such cases.

5.2 Gospels

For evaluating our algorithm's accuracy on the gospels, we again have to contend with the fact that the correct alignments are given at the resolution of pericopes, not verses. We cannot rely on the reciprocal best hit method we used for training, since it relies on the TF*IDF similarity scores, which we are attempting to evaluate. We therefore devise an alternative evaluation criterion, counting

[7]We discount the minor difference as insignificant.

a pair of verses as correctly aligned if they belong to a matched pericope in the gold annotation.

Let $Gold(g_1, g_2)$ be the set of matched pericope pairs for gospels g_1, g_2, according to Aland (1985). For each pair of matched verses, v_{g_1}, v_{g_2}, we count the pair as a true positive if and only if there is a pericope pair $\langle p_{g_1}, p_{g_2} \rangle \in Gold(g_1, g_2)$ such that $v_{g_i} \in p_{g_i}$, $i = 1, 2$. Otherwise, it is a false positive. Precision is defined as usual ($P = tp/(tp + fp)$).

For recall, we note that not all the verses of a matched pericope should be matched, especially when one pericope has substantially more verses than the other. In general, we may expect the number of verses to be matched to be the minimum of $| p_{g_1} |$ and $| p_{g_2} |$. We thus define recall as:

$$R = tp/\left(\sum_{\langle p_{g_1}, p_{g_2} \rangle \in Gold(g_1, g_2)} \min(| p_{g_1} |, | p_{g_2} |) \right) .$$

The results are given in Figure 6, including the word-overlap baseline, TF*IDF ranking with logistic regression, and the added global alignment. Once again, TF*IDF yields a substantial improvement over the baseline, and results are further improved by adding the global alignment.

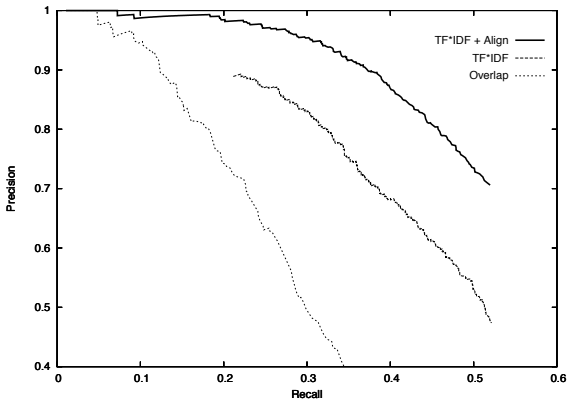

Figure 6: Precision/Recall curves for the gospels

6 Conclusions and future work

For monolingual alignment to achieve its full potential for text rewriting, huge amounts of text would need to be accurately aligned. Since monolingual corpora are so noisy, simple but effective methods as described in this paper will be required to ensure scalability.

We have presented a novel algorithm for aligning the sentences of monolingual corpora of comparable documents. Our algorithm not only yields substantially improved accuracy, but is also simpler and more robust than previous approaches. The efficacy of TF*IDF ranking is remarkable in the face of previous results. In particular, TF*IDF was not chosen by the feature selection algorithm of Hatzivassiloglou et al. (2001), who directly experimented and rejected TF*IDF measures as being less effective in determining similarity. We believe this striking difference can be attributed to the source of the weights. Recall that our TF*IDF weights treat each sentence as a separate document for the purpose of weighting. TF*IDF scores used in previous work are likely to have been obtained either by aggregation over the full document corpus, or by comparison with an external general collection, which is bound to yield lower discriminative power. To illustrate this, consider two words, such as the name of a city, and the name of a building in that city. Viewed globally, both words are likely to belong to the long tail of the Zipf distribution, having almost indistinguishable logarithmic IDF. However, in the encyclopedia entry describing the city, the city's name is likely to appear in many sentences, while the building name may appear only in the single sentence that refers to it, and thus the latter should be scored higher. Conversely, a word that is relatively frequent in general usage, e.g., "river" might be highly discriminative between sentences.

We further improve on the TF*IDF results by using a global alignment algorithm. We expect that more sophisticated sequence alignment techniques, as studied for biological sequence analysis, might yield improved results, in particular for comparing loosely matched document pairs involving non-linear text transformations such as inversions and translocations. Such methods could still modularly rely on the TF*IDF scoring.

We reiterate Barzilay and Elhadad's conclusion about the effectiveness of using the document context for the alignment of text. In fact, we are able to take better advantage of the intra-document context, while not relying on any assumptions about inter-document context that might be specific to one particular corpus. Identifying scalable principles for the use of inter-document context poses a challenging topic for future research.

We have restricted our attention here to pre-annotated corpora, allowing better comparison with previous work, and sidestepping the labor-intensive task of human annotation. Having es-

tablished a simple and robust document alignment method, we leave its application to much larger-scale document sets for future work.

Acknowledgments

We thank Regina Barzilay and Noemie Elhadad for providing access to the annotated Britannica corpus, and for discussion. This work was supported in part by National Science Foundation grant BCS-0236592.

References

Kurt Aland, editor. 1985. *Synopsis Quattuor Evangeliorum*. American Bible Society, 13th edition, December.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 1–8, April.

Pernilla Danielsson and Daniel Ridings. 1997. Practical presentation of a vanilla aligner. Research reports from the Department of Swedish, Goeteborg University GU-ISS-97-2, Sprakdata, February.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, Geneva, Switzerland, August.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

William A. Gale and Kenneth Ward Church. 1991. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184.

Vasileios Hatzivassiloglou, Judith L. Klavans, and Eleazar Eskin. 1999. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, College Park, Maryland.

Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. 2001. SIMFINDER: A flexible clustering tool for summarization. In *Proceedings of the Workshop on Automatic Summarization, pages 41–49. Association for Computational Linguistics, 2001.*

Hongyan Jing. 2002. Using hidden Markov modeling to decompose human-written summaries. *Computational Linguistics*, 28(4):527–543.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization – step one: Sentence compression. In *Proceedings of the American Association for Artificial Intelligence conference (AAAI)*.

Beth Levin. 1993. *English Verb Classes And Alternations: A Preliminary Investigation*. The University of Chicago Press.

Daniel Marcu. 1999. The automatic construction of large-scale corpora for summarization research. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 137–144. ACM.

Christof Monz and Maarten de Rijke. 2001. Lightweight subsumption checking for computational semantics. In Patrick Blackburn and Michael Kohlhase, editors, *Proceedings of the 3rd Workshop on Inference in Computational Semantics (ICoS-3)*, pages 59–72.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In Stephen D. Richardson, editor, *AMTA*, volume 2499 of *Lecture Notes in Computer Science*, pages 135–144. Springer.

Arcady R. Mushegian and Eugene V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academies of Science*, 93:10268–10273, September.

S.B. Needleman and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Chris Quirk, Chris Brockett, and William B. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149, Barcelona Spain, July.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Karen Spärck-Jones. 1972. Exhaustivity and specificity. *Journal of Documentation*, 28(1):11–21.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.