Title Generation for Machine-Translated Documents

Rong Jin
Language Technologies Institute
Carnegie Mellon University,
Pittsburgh, PA, 15213
Rong+@cs.cmu.edu

Alexander G. Hauptmann
Dept. of Computer Science
Carnegie Mellon University,
Pittsburgh, PA, 15213
Alex+@cs.cmu.edu

Abstract

In this paper, we present and compare automatically generated titles for machine-translated documents using several different statistics-based methods. A Naïve Bayesian, a K-Nearest Neighbour, a TF-IDF and an iterative Expectation-Maximization method for title generation were applied to 1000 original English news documents and again to the same documents translated from English into Portuguese, French or German and back to English using SYSTRAN. The AutoSummarization function of Microsoft Word was used as a base line. Results on several metrics show that the statistics-based methods of title generation for machine-translated documents are fairly language independent and title generation is possible at a level approaching the accuracy of titles generated for the original English documents.

1. Introduction

Before we discuss generating target language titles for documents in a different source language, let's consider in general the complex task of creating a title for a document: One has to understand what the document is about, one has to know what is characteristic of this document with respect to other documents, one has to know how a good title sounds to catch attention and how to distill the essence of the document into a title of just a few words. To generate a title for a machine-translated document becomes even more challenging because we have to deal with syntactic and semantic translation errors generated by the automatic translation system.

Generating text titles for machine translated documents is very worthwhile because it produces a very compact target language representation of the original foreign language document, which will help readers to understand the important information contained in the document quickly, without requiring a translation of the complete document. From the viewpoint of machine learning, studies on how well general title generation methods can be adapted to errorful machine translated documents and which methods perform better than others will be very helpful towards general understanding on how to discover knowledge from errorful or corrupted data and how to apply learned knowledge in 'noisy' environments.

Historically, the task of title generation is strongly connected to more traditional document summarization tasks (Goldstein *et al*, 1999) because title generation can be thought of as extremely short summarization. Traditional summarization has emphasized the extractive approach, using selected sentences or paragraphs from a document to provide a summary (Strzalkowski *et al*., 1998, Salton *et al*, 1997, Mitra *et al*, 1997). Most notably, McKeown *et al*. (1995) have developed systems that extract phrases and recombine elements of phrases into titles.

More recently, some researchers have moved toward "learning approaches" that take advantage of training data (Witbrock and Mittal, 1999). Their key idea was to estimate the probability of generating a particular title word given a word in the document. In their approach, they ignore all document words that do not appear in the title. Only document words that effectively *reappear* in the title of a document are counted when they estimate the probability of generating a title word wt given a document word wd as: P(wt|wd) where wt = wd. While the Witbrock/Mittal Naïve Bayesian approach is not in principle limited to this constraint, our experiments show that it is a very useful restriction. Kennedy and Hauptmann (2000) explored a generative approach with an iterative Expectation-Maximization algorithm using most of the document vocabulary. Jin and Hauptmann (2000a) extended this research with a comparison of several statistics-based title word selection methods.

In our approach to the title generation problem we will assume the following:
First, the system will be given a set of target language training data. Each datum consists of a document and its corresponding title. After exposure to the training corpus, the system should be able to generate a title for any unseen document. All source language documents will be translated into the target language before any title generation is attempted.

We decompose the title generation problem into two phases:

- **learning and analysis** of the training corpus and
- **generating a sequence of words** using learned statistics to form the title for a new document.

For **learning and analysis** of the training corpus, we present five different learning methods for comparison. All the approaches are described in detail in section 2.

- A Naïve Bayesian approach with limited vocabulary. This closely mirrors the experiments reported by Witbrock and Mittal (1999).
- A Naïve Bayesian approach with full vocabulary. Here, we compute the probability of generating a title word given a document word for all words in the training data, not just those document words that reappear on the titles.
- A KNN (k nearest neighbors) approach, which treats title generation as a special classification problem. We consider the titles in the training corpus as a fixed set of labels, and the task of generating a title for a new document is essentially the same as selecting an appropriate label (i.e. title) from the fixed set of training labels. The task reduces to finding the document in the training corpus, which is most similar to the current document to be titled. Standard document similarity vectors can be used. The new document title will be set to the title for the training document most similar to the current document.
- Iterative Expectation-Maximization approach. This duplicates the experiments reported by Kennedy and Hauptmann (2000).
- Term frequency and inverse document frequency (TFIDF) method (Salton and Buckley, 1988). TFIDF is a popular method used by the Information Retrieval community for measuring the importance of a term related to a document. We use the TFIDF score of a word as a measurement of the potential that this document word will be adopted into the title, since important words have higher chance of being used in the title.

For the **title-generating** phase, we can further decompose the issues involved as follows:

- Choosing appropriate title words. This is done by applying the learned knowledge from one of the six methods examined in this paper.
- Deciding how many title words are appropriate for this document title. In our experiments we simply fixed the length of generated titles to the average expected title length for all comparisons.
- Finding the correct sequence of title words that forms a readable title 'sentence'. This was done by applying a language model of title word trigrams to order the newly generated title word candidates into a linear sequence.

Finally, as a baseline, we also used the extractive summarization approach implemented as *AutoSummarize* in Microsoft Word that selects the "best" sentence from the document as a title.

The outline of this paper is as follows: This section gave an introduction to the title generation problem. Details of our approach and experiments are presented in Section 2. The results and their analysis are presented in Section 3. Section 4 discusses our conclusions drawn from the experiment and suggests possible improvements.

2. Title Generation Experiment across Multiple Languages.

We will first describe the data used in our experiments, and then explain and justify our evaluation metrics. The six learning approaches will then be described in more detail at the end of this section.

2.1 Data Description

The experimental dataset comes from a CD of 1997 broadcast news transcriptions published by Primary Source Media (1997). There were a total of roughly 50,000 news documents and corresponding titles in the dataset. The training dataset was formed by randomly picking four documents-title pairs from every five pairs in the original dataset. The size of training corpus was therefore 40,000 documents and their titles. We fixed the size of the test collection at 1,000 items from the unused document-title pairs. By separating training data and test data in this way, we ensure strong overlap in topic coverage between the training and test datasets, which gives the learning algorithms a chance to play a bigger role.

Since we did not have a large set of documents with titles in multiple parallel languages, we approximated this by creating machine-translated documents from the original 1000 training documents with titles as follows:
Each of the 1000 test documents was submitted to the SYSTRAN machine translation system (http://babelfish.altavista.com) and translated into French. The French translation was again submitted to the SYSTRAN translation system and translated back into English. This final retranslation resulted in our French machine translation data. The procedure was repeated on the original documents for translation into Portuguese and German to obtain two more machine translated sets of identical documents. For all languages, the average vocabulary overlap between the translated documents and the original documents was around 70%. This implies that at least 30% of the English words were translated incorrectly during translation from English to a foreign language and back to English.

2.2 Evaluation Metric

In this paper, two evaluations are used, one to measure **selection quality** and another to measure **accuracy of the sequential ordering** of the title words.

To measure the **selection quality** of title words, an F1 metric was used (Van Rjiesbergen, 1979). For an automatically generated title T_{auto}, F1 is measured against the correspondent human assigned title T_{human} as follows:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Precision and recall is measured as the number of identical words in T_{auto} and T_{human} over the number of words in T_{auto} and the number of words in T_{human} respectively. Since we are focusing here on the choice of appropriate title words, F1 is the most appropriate measure for this purpose. Obviously the sequential word order of the generated title words is ignored by this metric. However, preliminary tests with human ratings for the automatically generated titles suggest a strong correlation between F1 and human quality judgments.

Accuracy of the sequential ordering: To measure how well a generated title compared to the original human generated title in terms of word order, we measured the number of correct title words in the hypothesis titles that were in the same order as in the reference titles using the dynamic alignment algorithm described by (Nye, 1984).

To make all approaches comparable (except MS Word AutoSummarize and KNN), only 6 title words were generated by each method, as 6 was the average number of title words in the training corpus. Since AutoSummarize in Microsoft Word selects a complete sentence from the test document as the title, the restriction to a title length of exactly 6 words often prevents AutoSummarize from producing any title sentence. Thus, we allow longer titles for AutoSummarize in Microsoft Word. The KNN method always uses the complete title of the document in the training corpus most similar to the test document as the title for the test document and thus the restriction of six words does not apply to titles generated by KNN. Since we wanted to emphasize content word accuracy, stop words were removed throughout the training and testing documents and titles.

2.3 Description of Title Generation Approaches

As we mentioned in the introduction, we compared five statistics-based title generation methods together with the baseline "extractive" approach. They were:

1. **Naïve Bayesian** approach with **limited vocabulary** (NBL). Essentially, this algorithm duplicates the work by Witbrock and Mittal (1999), which tries to capture the correlation between the words in the document and the words in the title. For each title word TW, it counts the occurrences of document word DW, if DW is the same as TW (i.e. DW = TW). To generate a title, we merely apply the statistics to the test documents for generating titles and select the top title words TW where P(TW | DW) is largest.
2. **Naïve Bayesian** approach **with full vocabulary** (NBF). The previous approach counts only the cases where the title word and the document word are the same. This restriction is based on the assumption that a document word is only able to generate a title word with same surface string. The constraint can be easily relaxed by counting all the document-word-title-word pairs and apply this full statistics on generating titles for the test documents. In all other respects, this approach is the same as the previous one.
3. **K nearest neighbor** approach (KNN). This algorithm is similar to the KNN algorithm applied to topic classification in (Yang *et al*, 1994). It treats the titles in the training corpus as a set of fixed labels. For each new document, instead of creating a new title, it tries to find an appropriate "label", which is equivalent to searching the training document set for the closest related document. This training document title is then used for the new document. In our experiment, we use SMART (Salton, 1971) to index our training documents and test documents with the weight schema "ATC". The similarity between documents is defined as the dot product between document vectors. The training document closest related to the test document is found by computing the similarity between the test document and each training document (K=1).
4. **Iterative Expectation-Maximization** approach (EM). This algorithm reproduces the work by Kennedy and Hauptmann (2000), which treats title generation as a translation problem. We view a document as written in a 'verbose' language and its corresponding title as written in a 'concise' language. The approach builds a translation model (Brown et al., 1990), between verbose and concise languages, based on the documents and titles in the training corpus and applies the learned translation model to generate titles for new documents. The essential difference between this approach and Naïve Bayesian approaches is that EM treats the title word generation probability given a document as the sum of the title word generation probability from all the document words while the Naïve Bayesian approach treats it as the product of the title word generation probability from all document words.
5. **Term frequency and inverse document frequency approach** (TF.IDF). Term frequency TF, i.e. the frequency of words occurring in a document, shows how important a word is inside a document. Inverse document frequency IDF, i.e. the log of the total number of documents divided by the number of documents containing this word, shows how rarely a term appears in the collection. The product of these two factors, i.e. TF.IDF, gives the importance of a term related to a document (Salton and Buckley,

1988). The highest-ranking TF.IDF document words are chosen for the title word candidates. The inverse document frequency for each term is computed based on how often it appears in the training corpus.

6. **Extractive summarization** approach (AUTO). We use the AutoSummarize function built into Microsoft Word as a demonstration of an extractive approach, which select the "best" sentence from the document as the title.

2.4 The Sequencing Process for Title Word Candidates

To generate an ordered, linear set of candidates, equivalent to what we would expect to read from left to right, we built a statistical trigram language model using the CMU-Cambridge Spoken Language Modeling toolkit (Clarkson and Rosenfeld, 1997) and the 40,000 titles in the training set. This language model was used to determine the most likely order of the title word candidates generated by the NBL, NBF, EM and TF.IDF methods. The KNN and AUTO generated titles were already in natural sequence.

3 Experimental Results and Discussion

To illustrate the quality of the results, we first show an example of machine-generated titles and then present quantitative results and their analysis.

3.1 Example

The following is an excerpt of a document translated from English to Portuguese and back to English together with its original, human-assigned English title. The translation shows all the typical characteristics and problems of machine-translated documents. The corresponding machine-generated titles are shown in Table 1.

Original Title: O.J. SIMPSON CIVIL TRIAL - CASE GOES TO JURY SOON

Document:

… THE CIVIL EXPERIMENTATION DE SIMPSON OF THE J WILL BE SOON IN THE HANDS OF THE JURY. THE LAWYERS FOR BOTH THE SIDES GIVE TO ITS ADDITIONS THE ADVANCED FOLLOWING WEEK AND THEN THE DELIBERATIONS START. THE JURY IN THE CIVIL EXPERIMENTATION IS JANE CLAYSON OF B C. HERE. DE SIMPSON OF THE J IS IS OF THE CUT TODAY WHEN THE JUDGE AND THE LAWYERS IN BOTH THE SIDES TO WORK FOR ARE OF INSTRUCTIONS JURY. THE PLAINTIFFS HAD FINISHED ITS CASE IT REBUTTAL WITH SOME HARMFUL TESTIMONY OF A CONNOISSEUR WHO OF THE PHOTOGRAPH OMS AUTHENTICATED THIRTY PICTURES RECENTLY DISCOVERED DE SIMPSON IT J THAT CONSUMES LOW SHOES THE SAME DE BRUNO MAGLI STYLE RARE THE ASSASSIN CONSUMED. THE THEORIES OF THE CONTAMINATION OF THE DEFENSE ARE CITAÇÕES ABSOLUTELY RIDICULOUS OF CITATIONS. THE ARGUMENTS DE F THEY ARE PROGRAMADOS STILL TO START IN TUESDAY. THIS JURY COULD START THE CASE AND START TO DELIBERATE IN THURSDAY. …

Method	Title
NBL	white house simpson civil case jury
NBF	Continuing coverage simpson civil trial president
EM	Continuing coverage simpson civil trial jury
AUTO	civil experimentation de the arguments de f
KNN	oj simpson civil trial
TF.IDF	simpson jury start de jane additions

Table 1 shows the machine-generated titles for a document translated from English to Portuguese and back to English.

3.2 Results and Discussion

We compared the machine-generated titles against reference titles using both F1 metrics and number of correct title words in the correct order. Figure 1 and 2 show the F1 scores and the average number of correct title words in the correct order for each method over both original documents and translated documents.

Performance of learning methods is relatively language independent. According to both the metrics of F1 and the average number of title words in the correct order, the performance for documents translated from English to French and Portuguese and back to English is quite similar for all six different methods. Performance for documents translated from English to German and back to English is somewhat worse than for French and Portuguese. This may be due to the underlying SYSTRAN translation system, or inherent in the inflectional and noun-compounding characteristics of German which distinguish it from the other languages examined here.

AutoSummarization in Microsoft Word performs poorly. In terms of F1 and the average number of title words in the correct word, AutoSummarization from Microsoft Word performs much worse than the methods KNN, TF.IDF, NBL and EM over either the original documents or translated documents. The only exception is that AutoSummarization appears to work fine for the original documents in terms of the average number of title words in the correct order. We believe this is due to the fact that AutoSummarization is allowed a much longer title length, which increases the chance to catch the right title word and improves the average number of title words in the correct order. This confirms other research, which found that extractive summarization will only work well if there is much redundancy in the data and the summarization is much greater than 10% of the

document size (Mittal *et al* 1999). Furthermore, the extraction-based approach is unable to take full advantage of the training corpus.

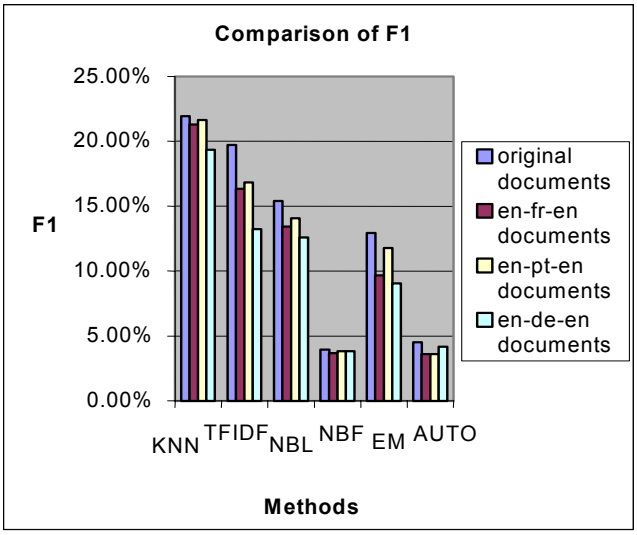

Figure 1 shows the F1 scores for the machine-generated titles. For each method, there are four bars representing the F1 scores for the titles generated from the original documents and the translated documents. The legends en-fr-en, en-pt-en and en-de-en represent the documents translated from English document to French, Portuguese and German and back in English, respectively.

K-Nearest Neighbor (KNN) performs extremely well. For both the original documents and the translated documents, KNN performs better than the other methods according to both the metrics of F1 and the average number of title words in the correct order. KNN works well here because the training and test sets were constructed to guarantee good overlap in content coverage. Even though our second best method, TF.IDF, shows performance close to KNN for the original documents, it degrades much more than KNN on all the three sets of machine-translated documents. There is almost no degradation for KNN over the documents translated from English to French and Portuguese and back to English. The large degradation for KNN over the document translated from English to German and back to English may be attributed to the fact that German is a quite different language from English as was already discussed. Actually, Jin and Hauptmann (2000b) have shown that KNN is also resilient to corruption from automatic speech recognition in generating titles for speech recognized documents. Thus, we conclude that KNN is a good approach for automatic title generation because of its simplicity and robustness to corrupt data.

Naive Bayesian with limit vocabulary (NBL) performs much better than Naïve Bayesian with full vocabulary (NBF). The difference between NBF and NBL is that NBL assumes a document word can only generate a title word with the same surface string. This very strong assumption discards information about a lot of words. However, the results tell us that some information can be safely ignored. In NBF, nothing distinguishes between important words and trivial words, and the co-occurrence between all document words and title words is measured equally. This lets frequent, but unimportant words dominate the document-word-title-word correlation. As an extreme example, stop words show up frequently in every document. However, they have little effect on choosing title words. Thus, even though NBF seems to exploit more knowledge than NBL, it introduces more noise by not limiting the effects of frequent, but unimportant words.

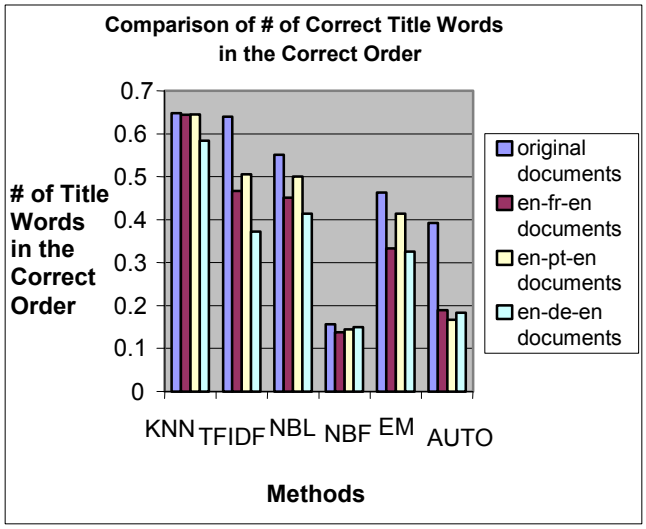

Figure 2 shows the average number of correct title words in the correct order. For each method, there are four bars representing the average number of correct title words in the correct order for the titles generated from the original documents and the translated documents. The legends en-fr-en, en-pt-en and en-de-en represent the documents translated from English document to French, Portuguese and German and back into English, respectively.

TF.IDF performs surprisingly well compared to the other true learning approaches. Surprisingly, the 'heavy' learning approaches, which take full advantage of the training corpus, such as Naïve Bayesian with limit vocabulary (NBL), Naïve Bayesian with full vocabulary (NBF) and Expectation-Maximization (EM) didn't outperform the shallow learning approach TF.IDF. Even though NBL, NBF and EM try to learn the association between title words and document words, they show no advantage over the simple TF.IDF approach, which selects the title words from the document based on the TF.IDF score, without learning anything about the titles. One suspicion is that learning the association between document words and title words by directly inspecting the document and its title is very problematic. Many words in a document don't reflect its content. For example, in many documents there are extraneous paragraphs and copyright notices. Those word occurrences will blur the statistics and mislead the title word selection. A better

strategy may be to first distill the document into essential content words and then compute the association between the distilled documents and their titles.

4 Conclusion

While title generation is far from a solved problem in one language, in this research we have, for the first time, applied learning approaches to title generation across languages. The research results show that automatic title generation is feasible on foreign language documents, despite gross errors in machine translation. Due to the flexibility and human readability issues of titles, the automatic evaluation metrics may not be able to reflect correctly the quality of titles. Thus, more work is needed to determine the human readability of the automatically generated titles, as well as the consistency between automatic evaluation metrics and human judgment. A validation with real cross-lingual documents is also desirable. In real life you would want to translate from French documents to English titles. You would train on English documents with English titles, and test on French documents translated into English.

References

[Goldstein *et al*, 1999] J. Goldstein, M. Kantrowitz, V. Mital, and J. Carbonell. Summarizing, Text Documents: Sentence Selection and Evaluation Metrics, *Proceedings of SIGIR 99*, Berkeley, CA, August 1999.

[Strzalkowski *et al*., 1998] T. Strzalkowski, J. Wang, and B. Wise, A robust practical text summarization system, *AAAI Intelligent Text Summarization Workshop*, pages 26--30, Stanford, CA, March 1998.

[Salton *et al*, 1997] G. Salton, A. Singhal, M. Mitra, and C. Buckley, Automatic text structuring and summary, *Info. Proc. And Management*, 33(2): 193-207, March 1997.

[Mitra *et al*, 1997] M. Mitra, A. Sighal, and C. Buckley, Automatic text summarization by paragraph extraction, *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain.

[McKeown *et al,* 1995] K. McKeown, J. Robin and K. Kukich, Generating Concise Natural Language Summaries, *Information Processing and Management*, 31 (5), pp.703-733, 1995.

[Witbrock and Mittal, 1999] M. Witbrock and V. Mittal, Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries, *Proceedings of SIGIR 99*, Berkeley, CA, August 1999

[Kennedy and Hauptmann, 2000] P. Kennedy and A.G. Hauptmann, Automatic Title Generation for the Informedia Multimedia Digital Library, *ACM Digital Libraries, DL-2000*, San Antonio Texas, May 2000, in press.

[Salton and Buckley, 1988] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, 24, 513—523, 1988

[Yang and Chute, 1994] Y. Yang and C.G. Chute, An example-based mapping method for text classification and retrieval, *ACM Transactions on Information Systems (TOIS)*, 12(3): 252-77. 1994.

[Van Rjiesbergen, 1979] Van Rjiesbergen. Butterworths, *Information Retrieval*, Chapter 7. London, 1979.

[Salton, 1971] G. Salton, *The SMART Retrival System: Experiments in Automatic Document Proceeding,* Prentice Hall, Englewood Cliffs, New Jersey. 1971.

[Clarkson and Rosenfeld, 1997] P.R. Clarkson and R. Rosenfeld. *Statistical Language Modeling Using the CMU-Cambridge Toolkit* Proceedings ESCA Eurospeech 1997

[Mittal, *et al,* 1999] V. Mittal, M. Kantrowitz, J. Goldstein and J. Carbonell, Selecting Text Spans for Document Summaries: Heuristics and Metrics, *AAAI-99*, 1999.

[Broadcast News, 1997] Primary Source Media, Broadcast News CDROM, Woodbridge, CT, 1997

[Nye, 1984] H. Nye, The Use of a One Stage Dynamic Programming Algorithm for Connected Word Recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. AASP-32, No 2, pp. 262-271, April 1984.

[Jin and Hauptmann, 2000a] R. Jin and A.G. Hauptmann, Cross Lingual Title Generation: Initial Steps, *Workshop on Interactive Searching in Foreign-Language Collections*, Human-Computer Interaction Laboratory, University of Maryland, College Park, MD. June 1, 2000. http://www.clis.umd.edu/conferences/hcil00

[Jin and Hauptmann, 2000b] R. Jin and A.G. Hauptmann, Title Generation for Spoken Broadcast News using a Training Corpus, In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, P.R.China, 2000

[Brown et al. 1990] P. Brown, S. Cocke, S. Della Pietra, Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and Roossin, A Statistical Approach to Machine Translation, Computational Linguistics V. 16, No. 2, June 1990.