Trans-EZ at NTCIR-2: Synset Co-occurrence Method for English-Chinese Cross-Lingual Information Retrieval

Guo-Wei Bian and Chi-Ching Lin

Trans-EZ Information Technology Inc.

14 F, No. 210 Sec. 1 Kee-Lung Rd., Taipei 110, Taiwan
URL: http://www.trans-ez.com
E-mail: arthur@trans-ez.com, cclin@trans-ez.com

Abstract

In this paper, a new method for English-Chinese cross-lingual information retrieval is proposed and evaluated in NTCIR-II project. We use the bilingual resources and contextual information to deal with the word sense disambiguation (WSD) and translation disambiguation for query translation. An English-Chinese WordNet and a synset co-occurrence model are adopted to solve the problem of word sense ambiguity. And the translation ambiguity and target polysemy are also resolved using such co-occurrence relationship of synsets. The experimental results are discussed to analyze the effects of ambiguity in source language and target language.

Keywords: *Synset, Word Sense Disambiguation, English-Chinese Information Retrieval.*

1 Introduction

Cross language information retrieval (CLIR) [13, 20, 21] deals with the use of queries in one language to access documents in another. Due to the differences between source and target languages, query translation is usually employed to unify the language in queries and documents. Some different approaches have been proposed for query translation. Dictionary-based approach exploits machine-readable dictionaries and selection strategies like select all [10, 16], randomly select N [1, 17] and select best N [10, 14]. Corpus-based approaches exploit sentence-aligned corpora [11] and document-aligned corpora [22]. These two approaches are complementary. Dictionary provides translation candidates, and corpus provides context to fit user intention. Coverage of dictionaries, alignment performance and domain shift of corpus are major problems of these two approaches. Hybrid approaches [2, 4, 10] integrate both lexical and corpus knowledge. A synset-based approach [8] is proposed to use an automatically constructed English-Chinese WordNet [7] for Chinese-English information retrieval.

Trans-EZ Information Technology Inc. has invested a lot of time and effort in Natural Language Processing (NLP) technology in Taiwan. We've been researching and developing several intelligent NLP-based systems, and integrating CLIR and MT together for a cross-language information access system [5]. In this system, users can express their information need and read the requested information in their familiar languages. Our previous paper [3] presents several important issues in an on-line and real-time document translation. Besides the translation ambiguity issue in query translation [4], we also touch on the target polysemy [6].

This paper will extend our works on Chinese-English CLIR and Japanese-English CLIR to English-Chinese CLIR. We use hybrid model, integrating dictionary-based and corpus-based approach, to resolve translation ambiguity problem. We employ dictionaries and co-occurrence statistics trained from target language documents to deal with translation ambiguity. This method considers the context around the translation equivalents to decide the best sense. The resources that we use are a bilingual dictionary, an English-Chinese WordNet, and a target language corpus. A bilingual dictionary provides the translation equivalents of each query term, and an English-Chinese WordNet provides the semantic synsets of words. The word co-occurrence information trained from target language text collection is used to disambiguate the senses of translation. The sense of a query term can be disambiguated using the co-occurrence of the senses of this term and other terms.

This paper is organized as follows. Section 2 shows the effects of ambiguities in Chinese-English and English-Chinese information retrievals. Section 3 presents the monolingual information retrieval in Chinese language. Section 4 deals with the English-Chinese cross-lingual information retrieval. Section 5 touches on the evaluation and discussion. Finally Section 6 concludes the remarks.

2 Effects of Ambiguities

Translation ambiguity and target polysemy are two major problems in CLIR [6]. Translation ambiguity results from the source language, and target polysemy occurs in target language. Take Chinese-English information retrieval (CEIR) and English-Chinese information retrieval (ECIR) as examples. The former uses Chinese queries to retrieve English documents, while the later employs English queries to retrieve Chinese documents. To explore the difficulties in the query translation of different languages, we gather the sense statistics of English and Chinese words. Table 1 shows the degree of word sense ambiguity (in terms of number of senses) in English and in Chinese, respectively. The Chinese thesaurus **同義詞詞林** (tong2yi4ci2ci2lin2) [19] and the English thesaurus Roget' s thesaurus are used to count the statistics of the senses of words. On the average, an English word has 1.687 senses, and a Chinese word has 1.397 senses. If the top 1000 high frequent words are considered, the English words have 3.527 senses, and the bi-character Chinese words only have 1.504 senses. For example, the Chinese word "**銀行**" (yin2hang2) is unambiguous, but its English translation "bank" has 9 senses [18]. In summary, Chinese word is comparatively unambiguous, so that translation ambiguity is not serious but target polysemy is serious in CEIR. In contrast, an English word is usually ambiguous. The translation disambiguation is important in ECIR.

Table 1. Statistics of Chinese and English Thesaurus

	Total Words	Average # of Senses	Average # of Senses for Top 1000 Words
English Thesaurus	29,380	1.687	3.527
Chinese Thesaurus	53,780	1.397	1.504

3 Monolingual IR

The test collection CIRB of NTCIR-II Chinese-IR Task is used to evaluate the performance of monolingual and cross-lingual information retrievals. This collection is composed of 50 topics (queries) in both English and Chinese to retrieve the Chinese document collection. Each query has relevance judgements. We use the Chinese topics to perform the Chinese monolingual retrieval and the English topics to perform the English-Chinese cross-language information retrieval.

In NTRIR-II CIRB collection, the original Chinese topics are composed of four fields: Title, Question, Narrative, and Concepts. In our experiments, only the fields of Title, Question, and Concepts are used to generate the queries. Because Chinese queries are composed of characters without word boundaries, the queries have to be segmented. A Chinese query is segmented by a word recognition system, and then tagged by a POS tagger. Our system selects the terms tagged with Noun or Verb as query terms. Regarding the document collection, we use our full-text search engine system to index the contents in Title and Text fields.

The topic CIRB010TopicZH001 is considered as an example in the following.

<topic>
<number>CIRB010TopicZH001</number>
<title>**集會遊行法與言論自由**</title>
<question>
查詢集會遊行法中有關主張共產主義或分裂國土規定之修正與討論。
</question>
<narrative>
相關文件內容應敘述集會遊行法原本對主張共產主義或分裂國土之限制，其是否符合憲法中對言論自由等基本人權的保障，大法官對此議題的相關解釋，學者專家的討論與看法，以及集會遊行法條文的修改現況。
</narrative>
<concepts>
集會遊行法、集會遊行、集遊法、憲法、言論自由、保障、共產主義、分裂國土、大法官會議、立法、修正條文。
</concepts>
</topic>

The query terms selected from the original Chinese query are listed below. The terms in the following lines are the results of the fields of Title, Question, and Concepts.

Title:	**'集會' '遊行' '法' '言論' '自由'**
Question:	**'集會' '遊行' '法' '主張' 共產主義' '分裂' '國土' '規定' '修正' '討論'**
Concepts:	**'集會' 遊行' 法' 集會' 遊行' 集遊法' '憲法' '言論' '自由' '保障' '共產主義' '分裂' 國土' 大法官' '會議' '立法' 修正' 條文'**

4 EC-CLIR

The recent works [2, 4] employ dictionaries and co-occurrence statistics trained from target language documents to deal with translation ambiguity. We will follow our previous work [4], which combines the dictionary-based and corpus-based approaches for CEIR. A bilingual dictionary provides the translation equivalents of each query term, and the word co-occurrence information trained from a target language text collection is used to disambiguate the translation. This method considers the context around the translation equivalents to decide the best target word. The translation of a query term can be disambiguated using the co-occurrence of the translation equivalents of this term and other terms. We adopt mutual information [9] to measure the strength. This disambiguation method performs good translations even when the multi-term phrases are not found in the bilingual dictionary, or the phrases are not identified in the source language. The recent work [8] adopts an English-Chinese WordNet to resolve the problem of translation ambiguity for Chinese-English information retrieval. We extend their method to use such bilingual WordNet-like resource to solve the problems of translation ambiguity and target polysemy in our English-Chinese information retrieval system. In a cross-language information retrieval system, the ambiguity of senses for a query term will grow from source language to target language during query translation. How to incorporate the knowledge from source side to target side is an important issue. We use the synset information in the English-Chinese WordNet to solve the word sense ambiguities in source language and target language.

4.1 Query Translation

For each English query, only the fields Title, Question, and Concepts are used to generate the queries. The stop words are filtered out in the queries also. The translation disambiguation of a query term is solved using the context information. The size of context for disambiguation processing will affect the correctness of translation and speed performance. The translation segments are decided using the punctuations ",", ".", "?", and "!".

We adopt a bilingual English-Chinese dictionary, an English-Chinese WordNet [7], and a Chinese corpus to solve the problem of translation ambiguity. The English-Chinese WordNet is used to solve the sense ambiguity both in the source language (English) and the target language (Chinese). Our system uses the first two resources to obtain the translation candiates and the sense candidates of query terms. Within each translation unit, we use the following Synset CO-Model to find the sense of each English query term. The Chinese terms in the selected synset can be regarded as the target query terms like the query expansion.

The processing of English query is shown in Figure 1.

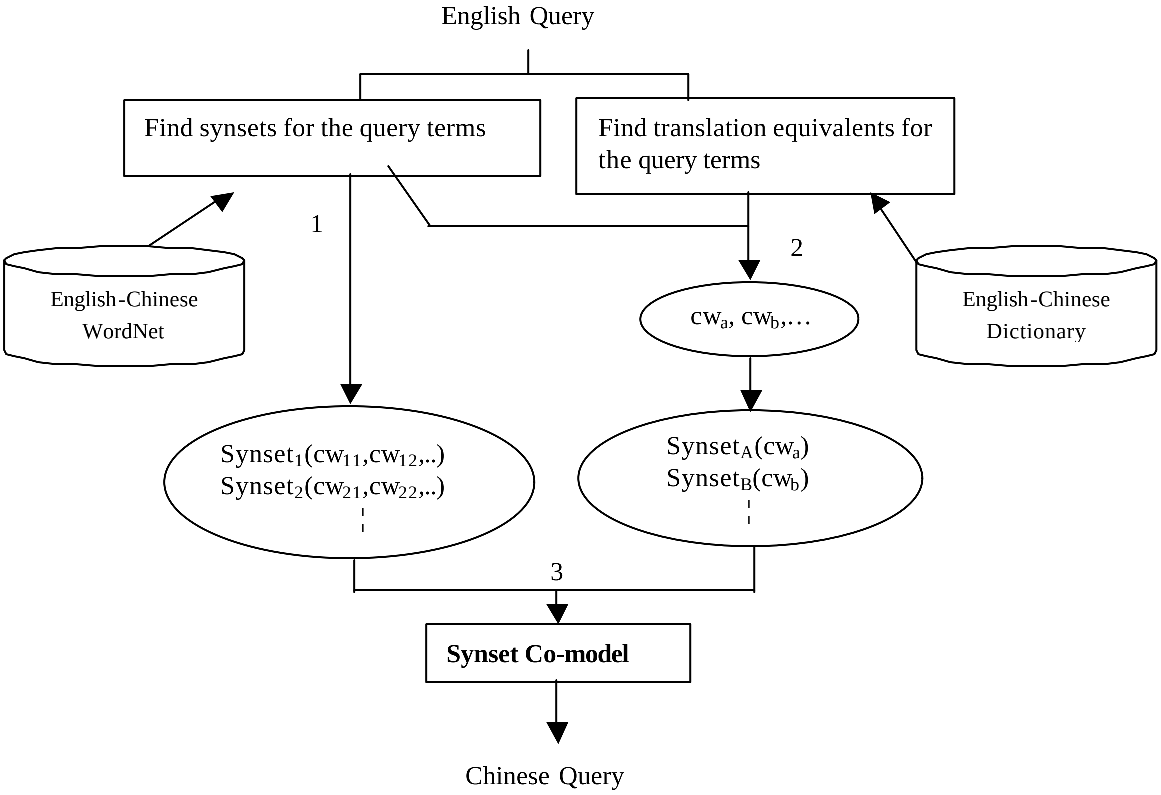
English Query
Find synsets for the query terms
Find translation equivalents for the query terms
1
2
English-Chinese WordNet
cw$_a$, cw$_b$,…
English-Chinese Dictionary
Synset$_1$(cw$_{11}$,cw$_{12}$,..)
Synset$_2$(cw$_{21}$,cw$_{22}$,..)
Synset$_A$(cw$_a$)
Synset$_B$(cw$_b$)
3
Synset Co-model
Chinese Query

Figure 1. The Processing Flow of Query Translation

1. After removing the stop words, we look up the English-Chinese WordNet for the remaining English words. A set of synsets is retrieved for each English query term. In the English-Chinese WordNet, a synset is composed of the English words like the WordNet. Additionally, a synset may contain some Chinese words. We eliminate those synsets with English words only, and the remaining bilingual synsets are used to solve the preoblems of translation ambiguity and target polysemy.
2. The Chinese equivalents of each English query term are obtained from the bilingual English-Chinese dictionary. For those Chinese equivalents not appearing in the synsets obtained in Step 1, we treat each of them as the individual synsets composing of one Chinese word only.
3. From the steps 1 and 2, each query term can be found out its synsets and the translation equivalents of synsets. We use the following Synset CO-Model to find the Chinese equivalents of each English query term.

The topic CIRB010TopicEN001 is considered as an example in the following.

```
<topic>
<number>CIRB010TopicEN001</number>
<title>The Assembly Parade Law and freedom of
     speech</title>
<question>
To retrieve the amendment and discussion of the
     regulations about communism and country
     separation in the Assembly Parade Law.
</question>
<narrative>
The content of related documents should be focused
     on the restrictions of communism and country
     separation originally proposed by the Assembly
     Parade Law, whether it conforms to constitution
     about freedom of speech indemnification, the
     interpretations of the grand justice regarding to
     this topic, discussions and viewpoints of
     experts/scholars, and the current status of the
     Assembly Parade Law amendments.
</narrative>
<concepts>
the Assembly Parade Law, Parade and Demonstration,
     constitution, freedom of speech,
     indemnification, communism, country
     separation, Council of Grand Justices,
     legislation, amendments.
</concepts>
</topic>
```

The synsets of English query term ‘Assembly’ in English-Chinese WordNet are listed below. The first column is the sense ID. of synset, and the others are the English terms and the Chinese terms in the synset.

Assembly

00585619_04_n	fabrication assembly
00798100_04_n	assembly assemblage gathering // 聚集
02217607_06_n	assembly // 組合
02716453_06_n	forum assembly meeting_place // 部件 論壇 配件 會場 集會 棋壇 影壇 球壇 乒壇
06071059_14_n	assembly

In the English-Chinese bilingual dictionary, the translation equivalents of the term ‘Assembly’ are:

Assembly 配件 組合 部件 集會

The synsets for each term in the query string “Assembly Parade Law” are listed in the following. The synset ID. 999999 indicated the special synset for those translation equivalents appearing in bilingual dictionary after the processing of step 2.

Assembly 配件 組合 部件 集會

02217607_06_n	組合
02716453_06_n	部件 論壇 配件 會場 集會 棋壇 影壇 球壇 乒壇
00798100_04_n	聚集

Parade 游行 遊行 標榜 閱兵

01313443_38_v	閱兵
01313330_38_v	散步 游行 漫步 信馬由韁 信步
999999	遊行
999999	標榜

Law 方法 定律 法 法制 法度 法律 法紀 法規 律 規率 綱紀 觚

04904589_10_n	定律 方法 律 法制
06093563_14_n	警察 法律 綱紀 法紀 巡警 法規 法度 巡捕 律 法制 紀綱
999999	法
999999	規率
999999	觚

4.2 Synset Co-model

We follow the strategy discussed previously for translation disambiguation [4]. This method considers the context around the English translation equivalents to decide the best target equivalent. Furthermore, an English-Chinese WordNet is used to solve the problem of translation ambiguity in English-Chinese information retrieval.

We compute the mutual information for the sets of synsets, and select a synset for each Chinese query term. Due to lack of sense-tagged coropa, we cannot compute the mutural information for the sysnset pairs directly. The mutual information of two synsets can be measured as follows. Let synset$_1$ and synset$_2$ be synsets for two query terms. Assume synset$_1$ and synset$_2$ are composed of *m* and *n* Chinese words, respectively.

$$MI\,(synset_1, synset_2) = \sum_{i=1}^{m}\sum_{j=1}^{n} MI(t_{1i}, t_{2j})\,/(m+n)$$

Where, the synset (synset$_1$) contains the words t$_{11}$, t$_{12}$, . .,.t$_{1m}$.
the synset (synset$_2$) contains the words t$_{21}$, t$_{22}$, . .,.t$_{2n}$.

The MI values of any two Chinese words are trained from ASBC corpus [15]. Chen, Lin, and Lin [8] proposed a method to measure the MI of synsets in Chinese-English information retrieval. We replace the (m*n) in the synset MI function as (m+n) to avoid preferring the smaller synset pairs. Table 2 shows the MI values of two synsets for the English query (Ew$_1$ Ew$_2$ Ew$_3$). We will select the synsets syn$_{11}$, syn$_{22}$, and syn$_{31}$ as the senses of the query term Ew$_1$, Ew$_2$, and Ew$_3$, respectively. In summary, our Synset-CO model employs the mutural information of words to select the appropriate synsets.

Table 2. MI Values of any Two Synsets in the Query

		Ew$_1$		Ew$_2$		Ew$_3$		
		syn$_{11}$	syn$_{12}$	syn$_{21}$	**syn$_{22}$**	**syn$_{31}$**	syn$_{32}$	syn$_{33}$
Ew$_1$	syn$_{11}$			1.517	**4.394**	1.233	0.444	1.583
	syn$_{12}$			—	—	—	—	—
Ew$_2$	syn$_{21}$	1.517	—			-0.061	0.028	-0.536
	syn$_{22}$	**4.394**	—			**3.899**	—	0.417
Ew$_3$	syn$_{31}$	1.233	—	-0.061	3.899			
	syn$_{32}$	0.444	—	0.028	—			
	syn$_{33}$	1.583	—	-0.536	0.417			

After step 3, the result of the query string “Assembly Parade Law” is in the following. In partically, the target terms for each query term are the words in the same synset.

Assembly	**部件 論壇 配件 會場 集會 棋壇 影壇 球壇 乒壇**
Parade	**遊行**
Law	**法**

Finally, the query string “Assembly Parade Law” is translated as the query “**部件 論壇 配件 會場 集會 棋壇 影壇 球壇 乒壇 遊行 法**” like query expansion does [12].

5 Experiments

The eleven-point average precision on the top 1,000 retrieved documents is adopted to measure the performance of all the experiments. The monolingual information retrieval, i.e., the original Chinese queries to Chinese text collection, is regarded as a baseline model. The performance is 0.3880 under the specified environment. In the English-Chinese experiment, the performance is 0.1318 only.

There are some factors to influence the cross-language retrieval performance. One is the coverage of the bilingual dictionary and the English-Chinese WordNet. Another one is the methodology to resolve the translation ambiguity. In our experiments, the English-Chinese WordNet is automatically constructed by mapping the Chinese words to the English WordNet. The correctness of such mapping is a problem to influence the correctness of senses. Another problem is whether the semantic structure of English WordNet is suitable to Chinese words. In the other hand, the methodology of synset co-occurrence model for the translation disambiguation has some problems to be solved. For those Chinese translation equivalents not appearing in the retrieval synsets of English query terms, we treat each of them as an pseudo synset composing of one Chinese word only. Such processing may make a mistake to split the Chinese translation equivalents with the same sense to different synsets. This problem may influence the selection of synsets. Further, the co-occurrence relationships of synsets cannot be measured directly because of the unavailability of sense-tagged corpora in Chinese. The measurement of proposed approach cannot reflect the association of synsets perfectly.

6 Conclusion

In this paper, we describe an English-Chinese cross-language information retrieval system for the evaluation in NTCIR-II ECIR task. We extend our work on Chinese-English CLIR to deal with this problem. We propose an approach to adopt the English-Chinese WordNet in CLIR system. The sense information is used to resolve the problems of translation ambiguity and expand the target query. The performance (0.1318) of English-Chinese information retrieval only achieves about 30% of the monolingual (Chinese) retrieval performance (0.3880). According to the degree of word sense ambiguity in English, the translation disambiguation is more serious than target polysemy in ECIR. Chen, Lin, and Lin [8] adopt the Chinese-English WordNet and propose different approaches in Chinese-English cross-language information retrieval. In their experiments on TREC-6 collection, the best performance is 0.1010. However, the experimental sets and the translation directions are totally different.

References

[1] Ballesteros, L. and Croft, W.B. (1996) “Dictionary-based Methods for Cross-Lingual Information Retrieval.” *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, 791-801.

[2] Ballesteros, L. and Croft, W.B. (1998) “Resolving Ambiguity for Cross-Language Retrieval.” *Proceedings of 21st ACM SIGIR*, 64-71.

[3] Bian, G.W. and Chen, H.H. (1997) “An MT Meta-Server for Information Retrieval on WWW.” In *Working Notes of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, Palo Alto, California, USA, March, 1997, pp.10-16.

[4] Bian, G.W. and Chen, H.H. (1998) “Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System.” *Machine Translation and Information Soup*, Lecture Notes in Computer Science, No. 1529, Spring-Verlag, 250-265.

[5] Bian, GW. and Chen, HH. (2000) “Cross language information access to multilingual collections on the Internet.” *Journal of American Society for Information Science*, 2000, **51**(3), pp. 281-296.

[6] Chen, H.H.; Bian, G.W.; and Lin, W.C. (1999) “Resolving translation ambiguity and target polysemy in cross-language information retrieval.” *Proceedings of 37th Annual Meeting of Association for Computational Linguistics*,1999, pp.215-222.

[7] Chen, H.H. and Lin, C.C. (2000) “Sense-Tagging Chinese Corpus.” *Proceddings of ACL Workshop on Chinese Language Processing*, 2000, pp. 7-14.

[8] Chen, H.H.; Lin, C.C.; and Lin, W.C. (2000) "Construction of a Chinese-English WordNet and Its Application to CLIR." *Proceedings of 5th International Workshop on Information Retrieval with Asian Languages*, September 30-October 2, Hong Kong, 189-196.

[9] Church, K. *et al*. (1989) “Parsing, Word Associations and Typical Predicate-Argument Relations.” *Proceedings of International Workshop on Parsing Technologies*, 389-398.

[10] Davis, M.W. (1997) “New Experiments in Cross-Language Text Retrieval at NMSU’s Computing Research Lab.” *Proceedings of TREC 5*, 39-1~39-19.

[11] Davis, M.W. and Dunning, T. (1996) “A TREC Evaluation of Query Translation Methods for Multi-lingual Text Retrieval.” *Proceedings of TREC-4*, 1996.

[12] Fitzpatrick, L. and Dent, M. (1997) “Automatic Feedback Using Past Queries: Social Searching. ” *Proceedings of 20th ACM SIGIR*, 306-313.

[13] Harman, D.K. (1997) *TREC-6 Proceedings*, Gaithersburg, Maryland.

[14] Hayashi, Y., Kikui, G. and Susaki, S. (1997) “TITAN: A Cross-linguistic Search Engine for the WWW.” *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*,58-65.

[15] Huang, C.R., *et al*. (1995) “Introduction to Academia Sinica Balanced Corpus. ” *Proceedings of ROCLING VIII*, Taiwan, 81-99.

[16] Hull, D.A. and Grefenstette, G. (1996) “Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval.” *Proceedings of the 19th ACM SIGIR*, 49-57.

[17] Kowk, K.L. (1997) “Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment.” *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, 110-114.

[18] Longman (1978) *Longman Dictionary of Contemporary English*. Longman Group Limited.

[19] Mei, J.; *et al*. (1982) *tong2yi4ci2ci2lin2*. Shanghai Dictionary Press.

[20] Oard, D.W. (1997) “Alternative Approaches for Cross-Language Text Retrieval.” *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, 131-139.

[21] Oard, D.W. and Dorr, B.J. (1996) *A Survey of Multilingual Text Retrieval*. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies. http://www.ee.umd.edu/medlab/filter/papers/mlir.ps.

[22] Sheridan, P. and Ballerini, J.P. (1996) “Experiments in Multilingual Information Retrieval Using the SPIDER System.” *Proceedings of the 19th ACM SIGIR*, 58-65.