Learning Bilingual Collocations by Word-Level Sorting

Masahiko Haruno Satoru Ikehara Takefumi Yamazaki
NTT Communication Science Labs.
1-2356 Take Yokosuka-Shi
Kanagawa 238-03, Japan
haruno@nttkb.ntt.jp ikehara@nttkb.ntt.jp yamazaki@nttkb.ntt.jp

Abstract

This paper proposes a new method for learning bilingual collocations from sentence-aligned parallel corpora. Our method comprises two steps: (1) extracting useful word chunks (n-grams) by word-level sorting and (2) constructing bilingual collocations by combining the word-chunks acquired in stage (1). We apply the method to a very challenging text pair: a stock market bulletin in Japanese and its abstract in English. Domain specific collocations are well captured even if they were not contained in the dictionaries of economic terms.

1 Introduction

In the field of machine translation, there is a growing interest in corpus-based approaches (Sato and Nagao, 1990; Dagan and Church, 1994; Matsumoto et al., 1993; Kumano and Hirakawa, 1994; Smadja et al., 1996). The main motivation behind this is to well handle domain specific expressions. Each application domain has various kinds of collocations ranging from word-level to sentence-level. The correct use of these collocations greatly influences the quality of output texts. Because such detailed collocations are difficult to hand-compile, the automatic extraction of bilingual collocations is needed.

A number of studies have attempted to extract bilingual collocations from parallel corpora. These studies can be classified into two directions. One is based on the full parsing techniques. (Matsumoto et al., 1993) proposed a method to find out phrase-level correspondences, while resolving syntactic ambiguities at the same time. Their methods determine phrase correspondences by using the phrase structures of the two languages and existing bilingual dictionaries. Unfortunately these approaches are promising only for the comparatively short sentences that can be analyzed by a CKY type parser.

The other direction for extracting bilingual collocations involves statistics. (Fung, 1995) acquired bilingual word correspondences without sentence alignment. Although these methods are robust and assume no information source, their outputs are just word-word correspondences. (Kupiec, 1993; Kumano and Hirakawa, 1994) extracted noun phrase (NP) correspondences from aligned parallel corpora. In (Kupiec, 1993), NPs in English and French texts are first extracted by a NP recognizer. Their correspondence probabilities are then gradually refined by using an EM-like iteration algorithm. (Kumano and Hirakawa, 1994) first extracted Japanese NPs in the same way, and combined statistics with a bilingual dictionary for MT to find out NP correspondences. Although their approaches attained high accuracy for the task considered, the most crucial knowledge for MT is more complex correspondences such as NP-VP correspondences and sentence-level correspondences. It seems difficult to extend these statistical methods to a broader range of collocations because they are specialized to NPs or single words.

(Smadja et al., 1996) proposed a general method to extract a broader range of collocations. They first extract English collocations using the Xtract system (Smadja, 1993), and then look for French counterparts. Their search strategy is an iterative combination of two elements. This is based on the intuitive idea that "if a set of words constitutes a collocation, its subset will also be correlated". Although this idea is correct, the iterative combination strategy generates a number of useless expressions. In fact, Xtract employs a robust English parser to filter out the wrong collocations which form more than half the candidates. In other languages such as Japanese, parser-based pruning cannot be used. Another drawback of their approach is that only the longest n-gram is adopted. That is, when 'Japan-US auto trade talks' is adopted as a collocation, 'Japan-US' cannot be recognized as a collocation though it is independently used very often.

In this paper, we propose an alternative method based on word-level sorting. Our method comprises two steps: (1) extracting useful word chunks (n-grams) by word-level sorting and (2) constructing bilingual collocations by combining the word-chunks acquired at stage (1). Given sentence-aligned texts in two languages(Haruno and Yamazaki, 1996), the first step detects useful word chunks by sorting and counting all uninterrupted word sequences in sentences. In this phase, we developed a new technique for extracting only useful chunks. The second step of the method evaluates the statistical similarity of the word chunks appearing in the corresponding sentences. Most of the fixed (uninterrupted) collocations are directly extracted from the word chunks. More flexible (interrupted) collocations are acquired level by level by iteratively combining the chunks. The proposed method, which uses effective word-level sorting, not only extracts fixed collocations with high precision, but also avoids the combinatorial explosion involved in searching flexible collocations. In addition, our method is robust and suitable for real-world applications because it only assumes part-of-speech taggers for both languages. Even if the part-of-speech taggers make errors in word segmentation, the errors can be recovered in the word chunk extraction stage.

2 Two Types of Japanese-English Collocations

In this section, we briefly classify the types of Japanese-English collocations by using the material in Table 1 as an example. These texts were derived from a stock market bulletin written in Japanese and its abstract written in English, which were distributed electrically via a computer network.

In Table 1, (東京外為 /*Tokyo Forex*), (日米自動車問題 /*auto talks between Japan and the U.S.*) and (控えて /*ahead of*) are Japanese-English collocations whose elements constitute uninterrupted word sequences. We call hereafter this type of collocation **fixed collocation**. Although fixed collocation seems trivial, more than half of all useful collocations belong to this class. Thus, it is important to extract fixed collocations with high precision. In contrast, (ドルは〜で取り引きを終えた /*The U.S. currency was quoted at* 〜) and (ドルは〜で取り引きを終えた /*The dollar stood* 〜)[1] are constructed from interrupted word sequences. We will call this type of collocation **flexible collocation**. From the viewpoint of machine learning, flexible collocations are much more difficult to learn because they involve the combination of elements. The points when extracting flexible collocations is how the number of combination (candidates) can be reduced.

Our learning method is twofold according to the collocation types. First, useful uninterrupted word chunks are extracted by the word-level sorting method. To find out fixed collocations, we evaluate stochastic similarity of the chunks. Next, we iteratively combin the chunks to extract flexible collocations.

3 Extracting Useful Chunks by Word-Level Sorting

3.1 Previous Research

With the availability of large corpora and memory devices, there is once again growing interest in extracting n-grams with large values of n. (Nagao and Mori, 1994) introduced an efficient method for calculating an arbitrary number of n-grams from large corpora. When the length of a text is l bytes, it occupies l consecutive bytes in memory as depicted in Figure 1. First, another table of size l is prepared, each field of which represents a pointer to a substring. A substring pointed to by the $(i-1)$th entry of the table constitutes a string existing from the ith character to the end of the text string. Next, to extract common substrings, the pointer table is sorted in alphabetic order. Two adjacent words in the pointer table are compared and the lengths of coincident prefix parts are counted(Gonnet et al., 1992).

For example, when *'auto talks between Japan and the U.S.'* and *'auto talks between Japan and China'* are two adjacent words, the number of coincidences is 29 as in *'auto talks between Japan and '*. The n-gram frequency table is constructed by counting the number of pointers which represent the same prefix parts. Although the method is efficient for large corpora, it involves large volume of fractional and unnecessary expressions. The reason for this is that the method does not consider the inter-relationships between the extracted strings. That is, the method generates redundant substrings which are subsumed by longer strings.

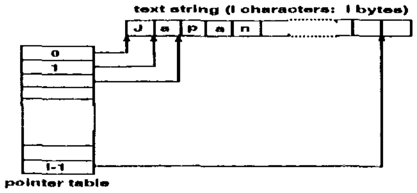


Figure 1: Nagao's Approach

To settle this problem, (Ikehara et al., 1996) proposed a method to extract only useful strings. Basically, his methods is based on the *longest-match* principle. When the method extracts a longest n-gram as a chunk, strings subsumed by the chunk are derived only if the shorter string often appears independently to the longest chunk. If *'auto talks between Japan and the U.S.'* is extracted as a chunk, *'Japan and the U.S.'* is also

[1] 〜 represents any sequence of words.

1. 東京外為１７時・円、小反発－－２６銭高８４円２１－２４銭。
Tokyo Forex 5 PM: Dollar at 84.21-84.24 yen

2. 前週末比２６銭円高ドル安の１ドル＝８４円２１－２４銭で大方の取引を終えた。
The dollar stood 0.26 yen lower at 84.21-84.24 at 5 p.m.

3. ドル安マルク高を受けて小口の円買いが先行したが、日米自動車問題での橋本通産相とカンター米通商代表部代表との会談を日本時間２７日未明に控えて模様眺めムードが強まり、円は８４円前半で小動きに推移した。
Forex market trading was extremely quiet ahead of further auto talks between Japan and the U.S., slated for early dawn Tuesday.

4. ドルは対マルクで反落し１ドル＝１．３８６３－６６マルクでほぼ取引を終えた。
The U.S. currency was quoted at 1.361-1.3863 German marks at 5:15 p.m.

Table 1: Sample of Target Texts

extracted because *'Japan and the U.S.'* is used so often independently as in *'Japan and the U.S. agreed ⋯'*. However, *'Japan and the'* is not extracted because it always appears in the context of *'Japan and the U.S.'*. The method strongly suppresses fractional and unnecessary expressions. More than 75 % of the strings extracted by Nagao's method are removed with the new method.

3.2 Word-Level Sorting Method

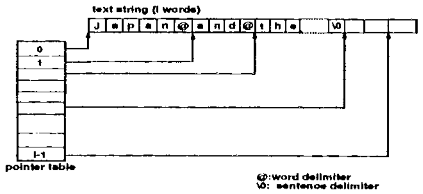
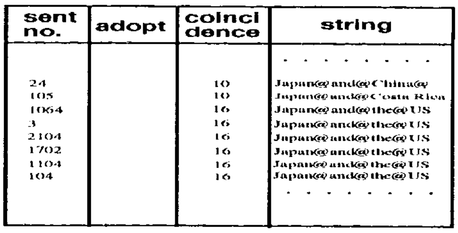


Figure 2: Word-Level Sorting Approach

The research described in the previous section deals with character-based n-grams, which generate excessive numbers of expressions and requires large memory for the pointer table. Thus, from a practical point of view, word-based n-grams are preferable in order to further suppress fractional expressions and pointer table use. In this paper, we extend Ikehara's method to handle word-based n-grams. First, both Japanese and English texts are part-of-speech (POS) tagged[2] and stored in memory as in Figure 2. POS tagging is required for two main reasons: (1) There are no explicit word delimiters in Japanese and (2) By using POS information, useless expressions can be removed.

In Figure 2, '@' and '\0' represent the explicit word delimiter and the explicit sentence delimiter, respectively. Compared to previous research, this data structure has the following advantages.

1. Only heads of each word are recorded in the pointer table. As depicted in Figure 2, this remarkably reduces memory use because the pointer table also contains other string characteristics as Figure 3.
2. As depicted in Figure 2, only expressions within a sentence are considered by introducing the explicit sentence delimiter '\0'.
3. Only word-level coincidences are extracted by introducing the explicit word delimiter '@'. This removes strings arising from a partial match of different words. For example, the coincident string between *'Japan and China'* and *'Japan and Costa Rica'* is *'Japan and'* in our method, while it is *'Japan and C'* in previous methods.



sent no.	adopt	coinci dence	string
			. . . .
24		10	Japan@and@China@
105		10	Japan@and@Costa Rica
1064		16	Japan@and@the@US
3		16	Japan@and@the@US
2104		16	Japan@and@the@US
1702		16	Japan@and@the@US
1104		16	Japan@and@the@US
104		16	Japan@and@the@US
			. . . .

Figure 3: Sorted Pointer Table

Next, the pointer table is sorted in alphabetic order as shown in Figure 3. In this table, **sentno.** and **coincidence** represent which sentence the string appeared in and how many characters are shared by the two adjacent strings, respectively. That is, **coincidence** delineates candidates for useful expressions. Note here that the coincidence between *Japan@and@China⋯* and *Japan@and@Costa Rica⋯* is 10 as mentioned above.

Next, in order to remove useless subsumed strings, the pointer table is sorted according to **sentno.**. In this stage, **adopt** is filled with '1' or '0' , each of which represents if or not if a string is subsumed by longer word chunks, respectively. Sorting by **sentno.** makes it much easier to check the subsumption of word chunks. When

[2] We use in this phase the JUMAN morphological analyzing system (Kurohashi et al., 1994) for tagging Japanese texts and Brill's transformation-based tagger (Brill, 1994) for tagging English texts. We would like to thank all people concerned for providing us with the tools.

both *'Japan and the U.S.'* and *'Japan and the'* arise from a sentence, the latter is removed because the former subsumes the latter.

Finally, to determine which word-chunks to extract, the pointer table is sorted once again in alphabetic order. In this stage, we count how many times a string whose **adopt** is 1 appears in the corpus. By thresholding the frequency, only useful word chunks are extracted.

4 Extracting Bilingual Collocations

In this section, we will explain how Japanese-English collocations are constructed from word chunks extracted in the previous stage. First, fixed collocations are induced in the following way. We use the contingency matrix to evaluate the similarity of word-chunk occurrences in both languages. Consider the contingency matrix, shown Table 2, for Japanese word chunk c_{jpn} and English word chunk c_{eng}. The contingency matrix shows: (a) the number of Japanese-English corresponding sentence pairs in which both c_{jpn} and c_{eng} were found, (b) the number of Japanese-English corresponding sentence pairs in which just c_{eng} was found, (c) the number of Japanese-English corresponding sentence pairs in which just c_{jpn} was found, (d) the number of Japanese-English corresponding sentence pairs in which neither chunk was found.

	c_{jpn}	
c_{eng}	a	b
	c	d

Table 2: Contingency Matrix

If c_{jpn} and c_{eng} are good translations of one another, a should be large, and b and c should be small. In contrast, if the two are not good translations of each other, a should be small, and b and c should be large. To make this argument more precise, we introduce mutual information as follows. Thresholding the mutual information extracts fixed collocations. Note that mutual information is reliable in this case because the frequency of each word chunk is thresholded at the word chunk extraction stage.

$$\log \frac{prob(c_{jpn}, c_{eng})}{prob(c_{jpn})prob(c_{eng})} = \log \frac{a(a+b+c+d)}{(a+b)(a+c)}$$

Next, we summarize how flexible collocations are extracted. The following is a series of procedures to extract flexible collocations.

1. For any pair of chunks in a Japanese sentence, compute mutual information. Combine the two chunks of highest mutual information. Iteratively repeat this procedure and construct a tree level by level.
2. For any pair of chunks in an English sentence, repeat the operations done in the the Japanese sentence.
3. Perform node matching between trees of both languages by using mutual information of Japanese and English word chunks.

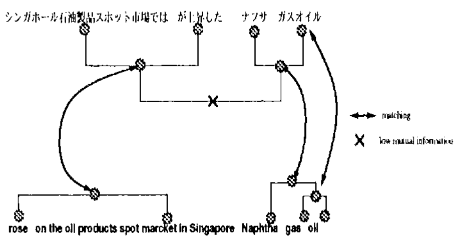

シンガポール石油製品スポット市場では
が上昇した
ナフサ
ガスオイル
matching
low mutual information
rose
on the oil products spot marcket in Singapore
Naphtha
gas
oil

Figure 4: Constructing Flexible Collocations

The first two steps construct monolingual similarity trees of word chunks in sentences. The third step iteratively evaluates the bilingual similarity of word chunk combinations by using the above trees. Consider the example below, in which the underlined word chunks construct a flexible collocation (シンガポール石油製品スポット市場では～が上昇した / ～ rose ～ on the oil products spot market in Singapore). First, two similarity trees are constructed as shown in Figure 4. Graph matching is then iteratively attempted by computing mutual information for groups of word chunks. In the present implementation, the system combines three word chunks at most. The technique we use is similar to the parsing-based methods for extracting bilingual collocation(Matsumoto et al., 1993). Our method replaces the parse trees with the similarity trees and thus avoids the combinatorial explosion inherent to the parsing-based methods.

Example:
シンガポール石油製品スポット市場では ナフサとガスオイル が上昇した
Naphtha and gas oil rose on the oil products spot market in Singapore

5 Preliminary Evaluation and Discussion

We performed a preliminary evaluation of the proposed method by using 10-days Japanese stock market bulletins and their English abstracts, each containing 2000 sentences. The text was first automatically aligned and then hand-checked by a human supervisor. A sample passage is displayed in Table 1.

In this experiment, we considered only the word chunks that appeared more than 4 times for fixed collocations and more than 6 times for flexible collocations. Table 4 illustrates the fixed collocations acquired by our method. Almost all collocations in Table 4 involve domain specific jargon, which

No.	Japanese	English
1	東京外為 〜 円 〜	Tokyo Forex ∼ Dollar at ∼ yen
2	ドルは 〜 で取り引きを終えた	The U.S. currency was quoted at ∼
3	〜 が売られ 〜 も安い	∼ were sold ∼ dropped as well
4	日銀が 〜 の供給を通知した	Bank of Japan injected ∼
5	オムロン 〜 住友林 〜	Omron ∼ Sumitomo Forestry ∼

Table 3: Samples of Flexible Collocations

No.	Japanese	English	No.	Japanese	English
1	東京外為	Tokyo Forex	38	債券相場	bonds and bond futures
2	控えて	ahead of	39	公的資金	public funds
3	マルク	German mark	40	機関投資家	institutional investors
4	ＪＡＦＣＯ	Japan Associated Finance	41	中心限月	benchmark
5	半面	in contrast	42	半導体関連株	semiconductor-related stocks
6	模様眺め気分が強い	remained sidelined watching	43	外国人投資家	foreign investors
7	警戒感	fear	44	ハイテク株	high-tech stocks
8	手控えられている	awaiting	45	売買高	turnover
9	東京貴金属大引け・	Tokyo Gold futures Cls:	46	小口の売り	small-lot selling
10	小動き	slow	47	更新した	record high
11	見送り気分	wait-and-see mood	48	指標銘柄	benchmark
12	アジアのディーラー間ドル建て相場	Loco-London gold	49	見送り気分が強く	low
13	対マルクで	against mark	50	東証２部	Tokyo Stocks 2nd Sec
14	ＣＢＱ平均	Convertible bonds	51	軟調	were weak
15	証券会社の自己売買部門	dealers	52	個人投資家	individual investors
16	売買高	trading volume	53	経常増益	pretax profit
17	利回り銘柄	high-yielders	54	第１部相場	The first section of TSE
18	日経３００先物・後場寄り	Nikkei 300 futures Aft-opg:	55	日経平均株価	the Nikkei stock average
19	前引け	mng-cls:	56	東証ＣＢ寄り付き・	Tokyo CBs Opg:
20	取り引きを終えた	contract ended	57	長期国債	long-term government bonds
21	緊急経済対策	economic stimulus package	58	で取引されている	were traded at
22	終り値は	closed at	59	輸入企業	importers
23	先物・大引け	futures cls:	60	は堅調	advanced
24	債券相場	bond market	61	買い戻し	covering
25	相場	convertible bonds	62	昭電工	Showa Denko
26	先物・後場寄り	nikkei futures aft-opg:	63	売買高は	volume was
27	嫌気し	disheartened by	64	年初来高値を更新し	hit a new year's high
28	への期待から	on expectation of	65	与党３党	ruling coalition
29	もしっかり	edged up	66	住特金	Sumitomo Special Metals
30	ハイテク株	high-tech shares	67	日経３００先物・前引け	Nikkei 300 futures Mng-cls:
31	様子見気分	wait-and-see mood	68	＜大証＞	OSE
32	住友林	Sumitomo Forestry	69	年初来安値	year's low
33	日米自動車交渉	U.S.-Japan auto talks	70	日経国際商品指数概況・	Nikkei World Commodities:
34	物色	speculative buying of	71	先物・寄り付き	nikkei futures opg:
35	東証外国部・前引け	Tokyo foreign stocks mng:	72	前引け	morning close
36	金利	interest rates	73	小動きに終始した	inched up
37	日米自動車交渉	the Japan-U.S. auto dispute			

Table 4: Samples of Fixed Collocations

cannot be constructed compositionally. For example, No 9 means "Tokyo Gold Future market ended trading for the day', but was never written as such. As well as No. 9 , a number of sentence-level collocations were also extracted. No. 9, No. 18, No. 23, No. 26, No. 35, No. 56 and No. 67 are typical heads of the stock market report. These expressions appear everyday in stock market reports.

It is interesting to notice the variety of fixed collocations. They differ in their constructions; noun phrases, verb phrases, prepositional phrases and sentence-level. Although conventional methods focus on noun phrases or try to encompass all kinds of collocations at the same time, we believe that fixed collocation is an important class of collocation. It is useful to intensively study fixed collocations because the collocation of more complex structures is difficult to learn regardless of the method used.

Table 3 exemplifies the flexible collocations we acquired from the same corpus. No. 1 to No. 4 are typical expressions in stock market reports. These collocation are extremely useful for template-based machine translation systems. No. 5 is an example of a useless collocation. Both Omron and Sumitomo Forestry are company names that co-occur frequently in stock market reports, but these two companies have no direct relation. In fact, more than half of all flexible collocations acquired were like No. 5. To remove useless collocations, constraints on the character types would be useful. Most useful Japanese flexible collocations contain at least one Hiragana[3] character. Thus,

[3] Japanese has three types of characters (Hiragana, Katakana, and Kanji), each of which has different amounts of information. In contrast, English has only

many useless collocations can be removed by imposing this constraint on extracted strings.

It is also interesting to compare our results with a Japanese-English dictionary for economics (Iwatsu, 1990). About half of Table 4 and all of Table 3 are not listed in the dictionary. In particular, no verb-phrase or sentence-level collocations are not covered. These collocations are more useful for translators than noun phrase collocations, but greatly differ from domain to domain. Thus, it is difficult in general to hand-compile a dictionary that contains these kinds of collocations. Because our method automatically extracts these collocations, it will be of significant use in compiling domain specific dictionaries.

Finally, we briefly describe the coverage of the proposed method. For the corpus examined, 70 % of the fixed collocations and 35 % of the flexible collocations output by the method were correct. This level of performance was achieved in the face of two problems.

- The English text was not a literal translation. Parts of Japanese sentence were often omitted and sometimes appeared in a different English sentence.
- The data set was too small.

We are now constructing a larger volume of corpus to address the second problem.

6 Conclusion

We have described a new method for learning bilingual collocations from parallel corpora. Our method consists of two steps: (1) extracting useful word chunks by the word-level sorting technique and (2) constructing bilingual collocations by combining these chunks. This architecture reflects the fact that fixed collocations play a more crucial role than accepted in previous research. Our method not only extracts fixed collocations with high precision but also reduces the combinatorial explosion that would be otherwise considered inescapable in extracting flexible collocations. Although our research is in the preliminary stage and tested with a small number of Japanese stock market bulletins and their English, the experimental results have shown a number of interesting collocations that are not contained in a dictionary of economic terms.

References

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. In *Proc. 12th AAAI*, pages 722–727.

Ido Dagan and Ken Church. 1994. *Termight:* identifying and translating technical terminology. In *Proc. Fourth Conference on Applied Natural Language Processing*, pages 34–40.

Pascale Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proc. 33rd ACL*, pages 236–243.

Gaston H. Gonnet, Ricardo A. Baeza-Yates, and Tim Snider, 1992. *Information Retrieval*, chapter 5, pages 66–82. Prentice-Hall.

Masahiko Haruno and Takefumi Yamazaki. 1996. High-Performance Bilingual Text Alignment Using Statistical and Dictionary Information. In *Proc. 34th ACL*.

Satoru Ikehara, Satoshi Shirai, and Hajime Uchino. 1996. A statistical method for extracting unitnerrupted and interrupted collocations from very large corpora. In *Proc. COLING96*.

Keisuke Iwatsu. 1990. *TREND: Japanese-English Dictionary of Current Terms*. Shougakkan.

Akira Kumano and Hideki Hirakawa. 1994. Building an MT dictionary from parallel texts based on linguisitic and statistical information. In *Proc. 15th COLING*, pages 76–81.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *the 31st Annual Meeting of ACL*, pages 17–22.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proc. International Workshop on Sharable Natural Language Resources*, pages 22–28.

Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. 1993. Structural matching of parallel texts. In *the 31st Annual Meeting of ACL*, pages 23–30.

Makoto Nagao and Shinsuke Mori. 1994. A new method of n-gram statistics for large number of n and automatic extraction of words and pharases from large text data of japanese,. In *Proc. 15th COLING*, pages 611–615.

Satoshi Sato and Makoto Nagao. 1990. Toward memory-based translation. In *Proc. 13th COLING*, pages 247–252.

Frank Smadja, Kathleen McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, March.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, March.

one type of character.