Words and Word Usage: Newspaper Text versus the Web

Vinci Liu and **James R. Curran**
School of Information Technologies
University of Sydney
NSW 2006, Australia
{vinci,james}@it.usyd.edu.au

Abstract

This paper explores the differences in words and word usage in two corpora – one derived from newspaper text and the other from the web. A corpus of web pages is compiled from a controlled traversal of the web, producing a topic-diverse collection of 2 billion words of web text[1]. We compare this *Web Corpus* with the Gigaword Corpus, a 2 billion word collection of news articles. The Web Corpus is applied to the task of *automatic thesaurus extraction*, obtaining similar overall results to using the Gigaword. The quality of synonyms extracted for each target word is dependent on the word's usage in the corpus. With many more words available on the web, a much larger Web Corpus can be created to obtain better results in different NLP tasks.

1 Introduction

In corpus-based Natural Language Processing (NLP), the corpus is the primary representation of a language from which algorithms extract information and build linguistic models. Words and word usage in a corpus of newspaper text will differ from that of a corpus of web pages, due to the different genres of text and as an artifact of the corpus collection process.

Words are used differently across different medium and corpora. While newspaper text is usually written by experienced writers and carefully checked by editors for accuracy, few restrictions exist for web text. It has a wider range of writing styles and less adherence to formal grammar. Anyone with access to the web can create web pages and there is no restriction as to what topics are written about on the web. Thus, a web corpus will contain a wider range of topics than a newspaper corpus.

Some differences between corpora can be attributed to the collection process. A corpus of newspaper text is usually collected from a few publishers across a set period of time. This corpus will reflect the topics in the news over that period as well as the types of news targeted by the publishers (e.g. political, financial). A corpus of web pages can contain pages from any time before its creation. It will have a different distribution of dates than a newspaper corpus.

In this paper, we explore the difference between a corpus of newspaper text and a corpus of web text. We conduct three experiments to highlight some of the differences. First, the token types that exist within each corpus are examined and the vocabulary unique to each corpus is accounted for. We then analysed words frequent in one corpus that are infrequent in the other to reveal topic skew. Finally we extract synonyms for common nouns to show the word usage and topic coverage of the two corpora and to demonstrate the usefulness of web corpora.

2 Corpora

A *corpus* is a collection of text fulfilling some specified criteria. If a corpus is intended to be used for study of English or any other language, it must incorporate samples across all usage of the target language. For example, an English corpus should include both written and spoken English. Each major type of text, across topic and genre, should be represented in the corpus in proportion to their usage in the language.

While a corpus can be broadly representative of a language, no collection of text can definitively represent a language. There is no set percentages that can be specified across mode and genre. We cannot ask how much more or less is English written than spoken? Instead, a corpus is better defined by its composition. In this paper, we compare two corpora – the LDC's Gigaword Corpus of newspaper text and our own Web Corpus.

[1] In this paper, we report on the number of tokens after the corpus has been tokenised, counting both words and punctuation.

Proceedings of the Australasian Language Technology Workshop 2005, pages 167–175,
Sydney, Australia, December 2005.

2.1 Existing Corpora

One of the first machine-readable corpora was the Brown Corpus (Francis and Kucera, 1979), created in 1964 and consisting of 1 million words of American English. Another step forward in corpus development came with the Penn Treebank, which consists of 4.5 million words of American English manually annotated with part of speech (POS) tags and parse trees (Marcus et al., 1994). The British National Corpus (BNC) is a collection of British English, consisting of 90 million words of written text and 10 million words of transcribed speech (Burnard, 2000). At almost one hundred times the size of the Brown Corpus and more than twenty times the size of the Penn Treebank, it is too large to be manually annotated and so the BNC is automatically tagged with POS tags.

While languages such as English are rich with language resources, minority languages often resort to using freely available web text. One of the first web-collected corpora was the Hungarian Web Corpus (Halacsy et al., 2004), created by downloading pages from the `.hu` domain. It has about 1 billion words of text after removal of duplicates and non-Hungarian text.

2.2 The Gigaword Corpus

The *English Gigaword Corpus* consists of over 4 million documents and 1.75 billion words (Graff, 2003), with more than 2 billion tokens when the text is tokenised, including punctuation. It is the next progression up in size from the BNC. The Gigaword is the large single collection of English news text available to-date. It consists of newspaper text from the Associated Press, Agence France Press (English Service), the New York Times Newswire, and Xinhua News Agency (English Service) from the years 1994-2001. Parts of Gigaword have been released by the LDC in other collections. The data is skewed toward the New York Times ($\sim$ 50%) and the Associated Press ($\sim$ 25%). The Agence France Press and Xinhua News Agency articles together make up the last 25%.

3 The Web Corpus

We collected the Web Corpus by a controlled traversal of the web. If a web spider were to traverse the web starting from a single seed URL, many more pages of the seed URL topics would be visited than other topics. As some topics on the web are linked to by a larger number of pages than others, these topics also tend to be over-represented in such a sample of the web. Pages pertaining to these topics tend to have more incoming links than others, but this is not entirely reflective of the popularity of such websites. Gambling and adult websites, for example, are known to densely link to one another.

3.1 Uniform Web Sampling

The web is too large to be downloaded entirely or for a significant percentage to be collected by most research projects. Two primary approaches exist for obtaining a uniform sample of the web. IP address sampling techniques (Lawrence and Giles, 1999; O'Neill et al., 1997) obtain a uniform sample by randomly generating addresses and exploring the associated server. While the IP address sampling approach has been successfully implemented and used for extraction statistics of the web, it is costly in the resources required. Lawrence and Giles report that only 1 in 269 tries of a random IP address received a response.

Random walk techniques (e.g. Henzinger et al., 2000) attempt to create a regular undirected web graph on which a random traversal would produce a uniform sample. This is usually accomplished using search engines to calculate the number of backward links (making the web undirected) and creating self-loops to standardise the number of links (both incoming and outgoing) for each page.

3.2 USyd-NLP-Spider

Our *Web Corpus* is compiled from the web using a method based on link-to-link traversal, similar to the random walk approaches. It allows faster download of web pages than the IP sampling technique but does not produce a uniform sample. Web pages are collected by the *USyd-NLP-Spider*, a multi-thread spider written in Python. We seeded the spider with links from the *Open Directory*[2]. The broad topic coverage of this open source classification tree allows us to create a topic-diverse collection of web text. However, certain topics in the directory have more links than others (not reflective of its coverage on the web) and topics of similar generality are placed at different depths. The Open Directory is flattened using a rule-based algorithm to reduce the topic skew. A list of 358 general topics and associated URLs is created.

From these seed URLs, the spider performs a breadth-first search. For each link, the spi-

[2] The Open Directory Project, `http://www.dmoz.org`

der samples pages from the same section of the website until a minimum word quota has been reached. External links are extracted and added to the link collection of the parent topic.

4 Text Cleaning

The HTML collected by the USyd-NLP-Spider must be transformed into a format usable by NLP algorithms – whitespace delimited tokens, organised into sentences, one per line. We call this process *text cleaning*. Text cleaning consists of many low-level processes, beginning with interpreting character encoding on HTML pages and transforming them into ISO Latin-1, followed by sentence boundary identification, tokenisation, and text filtering.

Our sentence boundary identification component is based on Ratnaparkhi (1998). We adapted his model for regular English text by adding additional features for HTML tags. Our tokeniser is based on the one used for the Penn Treebank (MacIntyre, 1995), modified to correctly tokenise URLs, email addresses, and other web-specific text.

The filtering component is especially important for cleaning web text. Not all parts of web pages consists of grammatical sentences; they may contain an ingredient list for a cooking recipe or fragment of C++ code. Our rule-based filter removes non-content words and foreign language text. It removes sentences and documents with a low percentage of dictionary words.

5 Token Types

We are interested in the type of tokens in each corpus. For example, are there more numbers on the web than in newspaper text? From each corpus, we randomly select a 1 billion word sample and classified the tokens into seven disjoint:

Numeric – At least one digit and zero or more punctuation characters, e.g. 2, 3.14, $5.50

Uppercase – Only uppercase, e.g. REUTERS

Title Case – An uppercase letter followed by one or more lowercase letters, e.g. Dilbert

Lowercase – Only lowercase, e.g. violin

Alphanumeric – At least one alphabetic and one digit (allowing for other characters), e.g. B2B, mp3, RedHat-9

Hyphenated Word – Alphabetic characters and hyphens, e.g. serb-dominated, vis-a-vis

Other – Any other tokens

Finally, we also measure the number of dictionary words using the Unix words file.

	Gigaword	Web Corpus
Numeric	1.8%	1.2%
Uppercase	1.4%	2.2%
Title Case	14.2%	14.4%
Lowercase	68.4%	68.7%
Alphanumeric	0.3%	0.2%
Hyphenated	0.9%	0.7%
Other	13.0%	12.6%
Dictionary Words	69.6%	66.9%

Table 1: Tokens for each corpus

	Gigaword		Web Corpus	
Tokens	1 billion		1 billion	
Token Types	2.2 million		4.8 million	
Numeric	343k	15.6%	374k	7.7%
Uppercase	95k	4.3%	241k	5.0%
Title Case	645k	29.3%	946k	19.6%
Lowercase	263k	12.0%	734k	15.2%
Alphanumeric	165k	7.6%	417k	8.6%
Hyphenated	533k	24.3%	970k	20.1%
Other	150k	6.8%	1,146k	23.7%
Dict. Words	43k	2.0%	45k	0.9%
% of Dict.	94.3%		98.0%	
45,427 words	42,835 words		44,539 words	

Table 2: Token types for each corpus

5.1 Token Classification

At the macroscopic level, the two corpora appear similar. Table 1 shows the percentage *by token* in each corpora across the seven categories. The results are very close, with the only significant difference being the 2.7% drop for dictionary words in the Web Corpus relative to the Gigaword. However, an analysis by *token type* shows big differences between the two corpora (see Table 2). The same size samples of the Gigaword and the Web Corpus have very different number of token types. While only 2.2 million token types are found in the 1 billion word sample of the Gigaword, about twice as many token types (4.8 million) are found in an equivalent sample of the Web Corpus.

An analysis of the token types show similar percentages in four of the seven categories: uppercase, lowercase, alphanumeric, and hyphenated tokens. Although the Web Corpus has about twice the number of *token types*, it has similar number of numeric token types as the Gigaword. The percentage of numeric token types in the Gigaword is more than twice that of the Web Corpus. The Web Corpus has a lower percentage of title case tokens, at 19.6%, than the Gigaword at 29.3%.

	Unique to Gigaword		Unique to Web Corpus	
All	1,413,427		4,048,531	
Numeric	282k	19.9%	313k	7.7%
Uppercase	36k	2.5%	182k	4.5%
Title Case	351k	24.8%	654k	16.2%
Lower Case	100k	7.1%	571k	14.1%
Alphanumeric	138k	9.8%	389k	9.6%
Hyphenated	395k	28.0%	832k	20.5%
Other	111k	7.9%	1,107k	27.3%
Dict. Words	0k	0.0%	2k	0.0%

Table 3: Token types unique to each corpus

A large percentage difference is also observed in the number of dictionary words. These percentages don't give the whole picture, as the Unix dictionary has only 45,427 words. Both corpora contain a high percentage of the words in the Unix dictionary, at 98.0% for the Web Corpus and 94.3% for the Gigaword.

The percentages of token types within a corpus is also very informative. While only 0.9% of the Web Corpus vocabulary is dictionary words, it accounts for 66.9% of the actual tokens. In the Gigaword, the dictionary words account for 2.0% of the token types but 69.6% of the token instances. About 734,000 (15.2%) of Web Corpus token types are lowercase, most of which are not found in the dictionary. Another 946,000 (19.6%) of Web Corpus token types are title case, which includes named entities. In the Web Corpus, and similarly in the Gigaword, the non-dictionary words are a large percentage of the token types but a relatively small percentage of the actual tokens.

5.2 Unique Token Types

To better account for the difference between the 2.2 million token types in the Gigaword compared with the 4.8 million token types in Web Corpus, we extracted the terms found in one corpus but not the other. Table 3 shows the percentage of token types unique to each corpus (i.e. found in the Gigaword but not in the Web Corpus, or vice-versa). Virtually no dictionary words are unique to each corpus, as both corpora already contain most of the words in the Unix dictionary.

Four significant categories are numeric, title case, hyphenated, and other tokens. They explain some of the difference between the vocabulary of the two corpora. Numeric tokens tend to be unique to texts; for example, the number 1,349,343 is unlikely to appear again in a different context. Title case tokens contains many named entities, which tend to be context specific. Hyphenated tokens behave more like bigrams as they are the combination of two unigrams. Other than the conventional hyphenated words (e.g. ice-cream), these bigram-like words tend to be more sparse. The above results suggest that the token types unique to Gigaword tend to be numbers and named-entities, whereas token types unique to the Web Corpus are non-standard words (e.g. email addresses and URLs).

5.3 Misspellings

A possible explanation for the significant difference between the number of token types is the misspelling of words. The web contains documents written by people with a widely varying command of English. Their work is not checked by professional editors unlike the newspaper text. Thus we expect that there are many more ungrammatical sentences and misspellings in the Web Corpus than the Gigaword. The misspellings in the Web Corpus are new "words" that contribute to the relatively higher token type count than the Gigaword.

To determine the degree that misspellings contribute to the number token types in the Web Corpus, we examined letter combinations that are one character away from the correct spelling. For a target word, we generate the letter combinations that are one operation from the correct spelling. Four operations are considered:

Insertion – A new letter is inserted into the correct word (not before the first letter)

Deletion – One letter in the correct word (except the first) is deleted

Substitution – One letter in the correct word (except the first) is substituted by another letter in the alphabet

Letter Reordering – One letter in the correct word (except the first) is swapped with the next letter

The only letter preserved in all of the above transformations is the first, as very few misspelling replaces the first letter of the word. Any combination found in a dictionary is also discounted, so that the correct word is not transformed into another valid word (e.g. difference to differences). Figure 4 shows the misspellings

Web Corpus		Gigaword
differeince	disfference	differencre
differrence	differience	differencce
differece	differenced	differnce
differenece	diffeerence	diffference
dfference	differenc	diference
differnce	differnence	diffrence
diffference	differennce	diffderence
diference	diffeence	differencel
diffrence		
3.7 matches per word		1.7 matches per word

Table 4: Misspelling of difference in Web Corpus and Gigaword

	Gigaword Rank	Web Corpus Rank	Diff. Rank
Kafelnikov	7,078	733,477	14
Vicario	9,658	613,056	19
Ivanisevic	7,147	569,627	23
Seles	5,285	179,175	77
McCurry	5,631	147,544	111
Walesa	7,287	146,494	112
Ciller	7,537	1,125,901	9
Serb-held	4,343	569,627	21
Muslim-Croat	8,791	381,462	32
SARAJEVO	9,556	300,220	38

Table 5: Selected words with Gigaword rank much higher than Web Corpus

of the word difference found in the Web Corpus and the Gigaword. While there are 17 misspellings of difference that are one transformation from the correct spelling in the Web Corpus, there are only 8 such misspellings in Gigaword. For all words found in the Unix dictionary, we calculated the average number of misspellings found in each of the two corpora. The Web Corpus has more than twice the number of misspellings than the Gigaword, 3.7 per word compared to 1.7 for the latter. Misspellings are another cause of the higher token type count for the Web Corpus.

6 Topical Words

Some topical differences between two corpora can be identified by finding words frequent in one corpora but not the other, and vice-versa. From each corpus we extract the 10,000 most frequent words and find the words with the biggest difference in rank between the corpora. This process highlights the differences between the two corpora, showing the words and topics with high coverage in one but little or no coverage in the other.

6.1 Frequent Gigaword Words

Table 5 shows examples of the top 10,000 ranked words in the Gigaword with the biggest difference with the Web Corpus rank. The words shown in the figure were selected to illustrate certain points and they are not indicative of all the words with a large difference in rank. The words can be divided into three groups:

The words in the first group, Kafelnikov, Vicario, Ivanisevic, and Seles, reflect the years covered by documents in the Gigaword. As the Gigaword contains newspaper articles from the years 1994-2001, these terms correspond to names of active professional tennis players of the time. This included Yevgeny Kafelnikov (active 1995-2004), Arantxa Sánchez Vicario (active 1989-2002), and Goran Ivanisevic (active 1988-2001). The Web Corpus on the other hand contains mostly texts from late 1990's onward, with a significant proportion written in the past few years. As these tennis players were no longer active (or no longer making the headlines) at the time that many Web Corpus documents were written, their names were not frequent terms in the Web Corpus.

The next two groups also reflect the news covered by the Gigaword articles. McCurry, Walesa, and Ciller are names of political figures during early and mid-1990's. Mike McCurry was the press secretary of U.S. President Bill Clinton from 1994-98, Lech Walesa was the Polish President from 1990-95, and Tansu Ciller was the Turkish Prime Minister from 1993-96.

The terms Serb-held, Muslim-Croat, and SARAJEVO in the third group are terms from newspaper articles about the Yugoslav War (a series of conflicts from 1991-2001). Possible phrases include Serb-held territories and Muslim-Croat army and SARAJEVO as the locational identifier at the start of an article.

6.2 Frequent Web Corpus Words

The terms dvd, MySQL, and mp3 were not found in the Gigaword. The all lowercase formatting of dvd and mp3 is likely the reason they were not found. While both were invented in the mid-1990's, they would probably always appear capitalised in newspapers text as DVD and MP3. MySQL, released in 1995, does not appear in the 1 billion word Gigaword sample.

Some web-oriented words with much higher ranks in the Web Corpus include unsubscribe and emailed. As the Internet only began to

	Web Corpus Rank	Gigaword Rank	Diff. Rank
dvd	6,546	Not found	16
MySQL	6,948	Not found	23
mp3	9,092	Not found	30
unsubscribe	8,932	753,428	47
emailed	8,102	641,461	52
pissing	8,337	351,980	63
pee	8,946	119,101	157

Table 6: Selected words with Web Corpus rank much higher than Gigaword

gain prominence only during the second half of the Gigaword timeline, such terms rarely appeared in that corpus. Many instances of the term unsubscribe may also have not been properly filtered out from the Web Corpus with non-content terms such as Click here to unsubscribe. This increased word rank of unsubscribe is an artifact of the text cleaning process of the Web Corpus.

Slang and expletives also have much lower usage in newspaper text. The terms pissing and pee, slang words for urinate, appear relatively more frequently in web text than in newspaper text. As newspaper text is carefully edited, use of expletives is restricted, and the use of slang and other colloquialisms is discouraged.

7 Thesaurus Extraction

Thesauri are useful in many NLP and Information Retrieval (IR) applications. They expand the recall and coverage of the system by providing synonyms of a target word. In NLP, for example, this expansion technique is helpful when n-gram counts for a target word are unreliable. In IR, synonyms help expand keyword queries into many related queries, boosting the recall rate of the system. While thesauri are traditionally manually collected, automatic thesaurus extraction is superior to manual construction in several aspects (Curran, 2004). Manual thesaurus construction is labour-intensive and time consuming, and the result suffers from bias, low coverage, and inconsistency. Bias and inconsistency of lexical resources can be seen in WordNet, in which similar categories of words have different degrees of distinction. As such lexical resources are constructed by human experts, their personal biases are also reflected in the final product. We extract thesauri from the Gigaword and the Web Corpus for the same set of headwords to see the differences in word usage and word similarity in each corpus.

7.1 Method

We used the thesaurus extraction system developed by Curran (2004). It is based on the *distributional hypothesis* that *similar words appear in similar contexts*. The system extracts one-word noun synonyms (i.e. not multi-word expressions). The extraction process is divided into two main parts. First, all target noun contexts are represented as relations and compiled into one context vector for each noun. Second, a comparison between all context vectors is made to identify the closest (i.e. most similar) terms.

Contexts are extracted from raw sentences using a maximum entropy POS tagger, chunker, and a relation extractor (Curran and Clark, 2003). Six different types of relationship are identified:

- Between a noun and a modifying adjective.
- Between a noun and a noun modifier.
- Between a verb and a subject.
- Between a verb and a direct object.
- Between a verb and an indirect object.
- Between a noun and the head of a modifying prepositional phrase.

The nouns in each case (including the subjects and objects) are the target headword. All context relations for a particular headword are aggregated into the headword's context vector. Words are identified as synonyms on the basis of the number of context vectors they have in common.

7.2 Evaluation

Curran evaluates against a combination of four gold standard thesauri: Macquarie (Bernard, 1990), Roget's (Roget, 1911), Moby (Ward, 1996), and Oxford (Hanks, 2000). The gold standard synonyms of a headword are aggregated into one unranked list. The *inverse rank* (InvR) evaluation metric takes the rankings within the extracted list into account. For example, if the extracted terms at ranks 3, 5, and 28 are found in the gold standard list, then InvR $= \frac{1}{3} + \frac{1}{5} + \frac{1}{28} \cong 0.569$.

200 synonyms are extracted for 300 headwords from 2 billion words of the Web Corpus and from 2 billion words of the Gigaword. The headwords are test nouns created to cover interesting properties – including across frequency bands of several corpora (Curran, 2004).

Corpus	InvR	InvR max
Gigaword	1.86	5.92
Web Corpus	1.81	5.92

Table 7: Average InvR for 300 headwords

	Word	InvR Scores			Diff.
1	picture	3.322	to	0.568	2.754
2	star	2.380	to	0.119	2.261
3	program	3.218	to	1.184	2.034
4	aristocrat	2.056	to	0.031	2.025
5	box	3.194	to	1.265	1.929
6	cent	2.389	to	0.503	1.886
7	home	2.306	to	0.523	1.783
8	newspaper	3.036	to	1.381	1.655
9	statement	3.199	to	1.629	1.570
10	firm	2.347	to	0.829	1.518

Table 8: Headwords with biggest InvR difference, Gigaword > Web Corpus

7.3 Results

Table 7 shows the average InvR scores for the Gigaword and the Web Corpus for the 300 headwords. While the overall performance of the two corpora are very similar, on a per word basis one corpus can significantly outperform the other.

7.4 Gigaword Higher InvR Score

Table 8 shows the top 10 terms which the Gigaword InvR results were better than Web Corpus. For the headword home, much better synonyms were extracted from the Gigaword. Table 9 shows the top 50 extracted terms from both corpora. A similar number of matches were made with the gold standard list, with 24 matches for Gigaword to 18 for the Web Corpus. However, the matches were among the top terms in Gigaword but not in the Web Corpus. The top two terms house and apartment were extracted from the Gigaword, but the terms such as page and loan were extracted from the Web Corpus. Collocations, such as home page, were incorrectly extracted instead of synonyms.

Gigaword (24 matches out of 200)
house apartment building run office resident **residence headquarters** victory **native place mansion** room trip mile **family** night hometown town win neighborhood life suburb school restaurant hotel store city street season area road homer day car shop **hospital** friend game farm facility center north child land weekend community loss **return** hour . . .
Web Corpus (18 matches out of 200)
page loan contact **house** us owner search finance mortgage office map links building faq equity news center estate privacy community info business car **site** web improvement extention heating rate directory room **apartment family** service rental credit shop life city school **property place location** job online vacation store facility library free . . .

Table 9: Synonyms for home

	Word	InvR Scores			Diff.
1	chain	3.139	to	0.224	2.915
2	walk	3.184	to	0.774	2.410
3	point	3.540	to	1.477	2.063
4	bloke	2.445	to	0.425	2.020
5	game	2.799	to	1.097	1.702
6	graph	2.400	to	0.714	1.686
7	reinforcement	1.808	to	0.244	1.564
8	announcement	1.993	to	0.495	1.498
9	sport	3.116	to	1.642	1.474
10	solicitor	1.634	to	0.161	1.473

Table 10: Headwords with biggest InvR difference, Web Corpus > Gigaword

7.5 Web Corpus Higher InvR Score

Table 10 shows the top 10 terms which the Web Corpus InvR results were better than Gigaword. The Web Corpus outperformed Gigaword in extracting synonyms for terms such as chain. Table 11 shows the top 50 extracted terms from both corpora. 53 gold standard synoyms were extracted out of the Web Corpus compared to only 9 for the Gigaword. This difference in performance can be attributed to the topic skew of the Gigaword and the gold standards. The terms extracted by Gigaword belong to only one sense of the word chain, as in chain stores. The gold standard terms included a more physical sense of chain, such as necklace chain.

A bias is apparent in the topic coverage of both Gigaword and the gold standard. Gigaword is skewed toward the business sense of chains, reflecting financial text that is a significant portion of newspaper articles. The gold standard is skewed toward other senses. The wide topic coverage of Web Corpus becomes apparent in this example. While the top extracted Web Corpus terms also corresponds to the physical sense of chains (e.g. necklace, bracelet, and pendant), terms were also extracted belonging to the business sense of the word (e.g. retailer). Synonyms of chain extracted from the Web Corpus have a much better coverage of the different senses of the word than Gigaword or the gold standard thesauri alone.

Gigaword (9 matches out of 200)
store retailer supermarket restaurant outlet operator shop shelf owner grocery **company** hotel manufacturer retail franchise clerk maker discount business sale superstore brand clothing food giant shopping **firm** retailing industry drugstore distributor supplier bar insurer inc. **conglomerate** network unit apparel boutique mall electronics carrier division brokerage toy producer pharmacy airline inc . . .
Web Corpus (53 matches out of 200)
necklace supply **bracelet pendant rope belt ring earring** gold bead silver **pin wire cord** reaction clasp jewelry **charm** frame **bangle strap** sterling loop timing plate metal **collar** turn **hook** arm length **string** retailer repair **strand** plug diamond wheel industry tube surface neck **brooch** store **molecule ribbon** pump **choker** shaft body . . .

Table 11: Synonyms for chain

Gigaword (13 out of 200)
acushnet zoolander working-class marshak interchangeability scouse ghyll dubliner **fella** film guy yorkshireman aussie bostonite irishman **lad** bumbler **chap** scrum-half texan ex-marine profane kansan medavoy **gentleman guy** ballplayer Irishman anybody lunk **somebody** up-and-down vaudevillian yorker theatricality englishman **person** hobby newspaperman klutz goof everyman chicagoan scotsman artilleryman brazilian **fellow** midwesterner ref ballclub . . .
Web Corpus (16 matches out of 200)
lad fella somebody bondsman endomorphism **gentleman** aussie **dude** boucher **guy** englishman **chap** stranger balfour iraqi youngster nobody policeman cop passer-by everybody waitress boyfriend anybody no-one **punter** mum irishman lowepro teenager businessman bartender girlfriend fiance buffy neighbour 40ml hippie **bastard** beggar sandstorm kiwi foreigner grandma frenchman dad yank pooch brit spectator . . .

Table 12: Synonyms for bloke

Web Corpus also significantly outperformed Gigaword for the term bloke (see Table 12). Bloke, British and Australian slang for a man, has a much higher InvR score on the Web Corpus list than the Gigaword list. This reflects the international nature of the web, where terms specific to British and Australian English were found often enough to be reliably characterised by their context. Documents included in Gigaword have a skew towards American English, with the New York Times contributing the majority of text in that corpora. Without many training examples in non-American English, it is difficult to correctly extract synonyms for words such as bloke.

7.6 Discussion

While the Gigaword and Web Corpus have similar overall averages in the InvR scores, there are significant differences in performance for different terms. The Gigaword consists of newspaper text and better synonyms are extracted for topics covered in the news. The Web Corpus is more international and more topic-diverse, successfully extracting synonyms in different varieties of English and for different senses of words. However the dominance of certain topics on the web, with web-specific vocabulary, means that sometimes a highly biased thesauri is extracted.

To create a better Web Corpus, not only is there a need to cover a wide ranging number of topics, but one must actively prevent specific topics from dominating the corpus. A more balanced corpus can be created with better spidering strategies. For example, the spider could be designed to automatically identify the topics of the websites visited.

8 Conclusion

The web is a promising source for creating large corpora for Natural Language Processing. In this paper, we compared our Web Corpus to the traditional Gigaword Corpus and demonstrated that the Web Corpus is useful for the task of automatic thesaurus creation.

Words and word usage differ in corpora, especially when they are compiled from different sources and medium. We examined the words and word usage in the Gigaword Corpus as compared with the Web Corpus. We have shown some of the differences in topics covered by the two corpora, as well as vocabulary variants and errors. Some of these contrasts can be attributed to the genre of text, but some are artifacts of the corpus creation process.

Our results in thesaurus extraction showed that the web text obtained similar overall results to a corpus of newspaper text. The alternative topical and lingustic information suggests that web-collected corpora is a viable addition or even alternative to traditional corpora of newspaper and other printed text.

As the Web Corpus is significantly larger than most corpora of printed text, better results can be obtained by training algorithms on the Web Corpus. This is especially true of tasks that suffer from the data sparseness problem. With much more text available for download on the web, the limits of the Web Corpus in size have yet to be reached.

Acknowledgements

We like to thank our anonymous reviewers and the Language Technology Research Group at the University of Sydney for their comments and feedback. This work has been supported by the Australian Research Council under Discovery Project DP0453131.

References

John R.L. Bernard, editor. 1990. *The Macquarie Encyclopedic Thesaurus.* The Macquarie Library, Sydney, Australia.

Lou Burnard, editor. 2000. *Reference Guide British National Corpus (World Edition).* British National Corpus Consortium.

James R. Curran and Stephen Clark. 2003. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 91–98, Budapest, Hungary, 12–17 April.

James Curran. 2004. *From Distributional to Semantic Similarity.* Ph.D. thesis, University of Edinburgh, Edinburgh, UK.

W. Nelson Francis and Henry Kucera. 1979. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Technical report, Brown University, Providence, RI USA.

David Graff. 2003. English Gigaword. Technical Report LDC2003T05, Linguistic Data Consortium, Philadelphia, PA USA.

Peter Halacsy, Andras Kornai, Laszlo Nemeth, Andras Rung, Istvan Szakadat, and Vikto Tron. 2004. Creating open language resources for Hungarian. In *Proceedings of the Language Resources and Evaluation (LREC)*, pages 203–210, Lisbon, Portugal.

Patrick Hanks, editor. 2000. *The New Oxford Thesaurus of English.* Oxford University Press, Oxford, UK.

M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. 2000. On Near-Uniform URL Sampling. In *Proceedings of the 9th International World Wide Web Conference.*

Steve Lawrence and C. Lee Giles. 1999. Accessibility of information on the web. *Nature*, 400:107–109, 8 July.

Robert MacIntyre. 1995. Sed script to produce Penn Treebank tokenization on arbitrary raw text. http://www.cis.upenn.edu/~treebank/tokenizer.sed.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Edward T. O'Neill, Patrick D. McClain, and Brain F. Lavoie. 1997. A Methodology for Sampling the World Wide Web. *Annual Review of Online Computer Library Center (OCLC) Research.*

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution.* Ph.D. thesis, University of Pennsylvania, Philadelphia, PA USA.

Peter Mark Roget. 1911. *Thesaurus of English words and phrases.* Longmans, Green and Company, London, UK. Available from http://promo.net/pg/.

Grady Ward. 1996. *Moby Thesaurus.* Moby Lexicon Project. Available from http://etext.icewire.com/moby/.