Automatic Acquisition of Lexical Knowledge from Sparse and Noisy Data

René Schneider

Daimler-Benz AG, Institute of Information Technology,
Department of Speech and Language Understanding, Ulm, Germany
rene.schneider@dbag.ulm.DaimlerBenz.COM

Abstract. Optical character recognition (OCR) still garbles a considerable amount of information reduction and noise on texts so that many documents are unsuitable for information extraction systems. This paper introduces a statistical method for bootstrapping a lexicon from a very small number of "noisy ," domain-specific texts. This method determines regularity in grammatical forms and also reoccuring ungrammatical forms from the input text. Through a combination of frequency lists and Levenshtein matrices, a language independent, robust core lexicon is constructed that supports the analysis of "noisy texts," too.

1 Motivation

The growth of electronically transmissible and freely available texts that has taken place in recent years has not lead to a reduction of paperbound text. Therefore, the development of optical character recognition (OCR) systems and the improvement of their efficiency is still a major task in the area of document processing. [1] But anyway, even with high quality scanners a 100% recognition rate remains the ideal case. Besides the mistakes caused by OCR, a considerable number of documents include typographical or grammatical mistakes (misspellings, wrong inflection or word order, etc.). Therefore new methodologies should be invented that enable NLP-Systems to learn automatically from very small and grammatically incorrect corpora.

Statistical learning algorithms are usually applied to processing large corpora, but in real life, huge samples are hard to find for commercial and industrial applications. In our case, the corpora consists of a small sample of short letters requesting annual business reports from a company.

2 Information Theoretical Background

2.1 Syntactic vs. Pragmatic Information

Traditionally [4], information was computed as the negative sum of the probabilities of certain events. In other words: the less frequent the event was, the higher its information value was. In the case of natural languages this means for example that the word *the* has a very low value of information compared

to very specific and less frequent words like e.g. *benign*. But what happens if the information is transmitted through a noisy channel? In this case the number of seldom and arbitrary sequences grows dramatically, e.g. sequences like *7O)ankyouverynouchhoadvance* (caused by the use of multifont) or even *+ -; p*mj -pL* (as a consequence of dirt specks).

In these cases Shannon and Weaver's measure of syntactic information is no longer applicable. A solution has to be found, regarding the other two dimensions of information, namely semantic and pragmatic information. Both are very hard to measure, due to their subjective nature and even semantic information presumes a noiseless channel. Therefore Weizsäcker [5] introduced the concept of *pragmatic information* that deals with the *where*, the *when*, and the *how* of information. In this way, information can only be computed after it has taken place and with respect to a given situation. Basically, pragmatic information consists of the complementary parts of *novelty* and *confirmation* . Novelty means that at a certain point in time an event occurs for the first time. But even when the information itself is new for us, generally we are able to make predictions about the speaker, the location, etc. The event includes at least something already known, quantified as confirmation. In other words, information exchange is only possible when sender and receiver have a common semantic basis. To give an

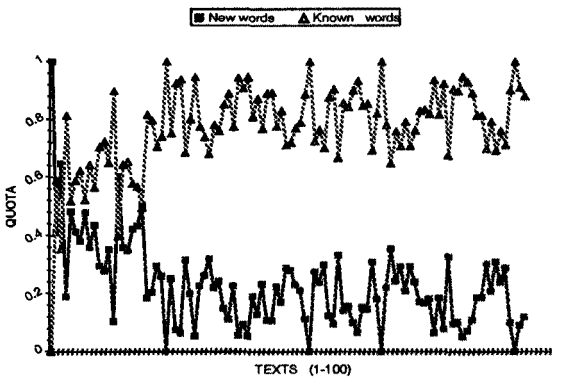

Fig. 1. Firstness and confirmation of word forms

example of the difference of novelty and confirmation, natural languages have a very high amount of novelty but require on the other hand constant confirmation, whereas artificial languages (e.g. machine code) constitute of nothing but confirmation and do not allow any kind of firstness or novelty.

2.2 Evaluation

Every text (or text body) of the training corpus can be seen as a closed unit and all the texts can be brought into a random order required to create a temporal structure. To better understand the concepts of novelty and confirmation, the relative amounts of unknown and verified words were computed. As can be seen

in Figure 1, the number of new words is very high at the beginning and rapidly decreases to a more or less constant value. The curve of confirmed words shows the opposite effect. After a very low number of texts, generally 80 % of the information is confirmed, i.e. the words appeared already in one of the former texts. These 80 % cover generally the functional words such as articles, conjunctions etc. and of course the domain–specific information. The residual 20 % consist of text–relevant information, unimportant and less interesting information, misspellings, and — in OCR–texts — noisy information. The curve's oscillation is caused by factors such as text size and the OCR quality of the different texts.

A linguistic interpretation of the different qualities of the words with a high confirmation value (see 3.1, Unordered lexicon entries) shows that they are usually correct and not inflected. These results lead to the following assumptions: 1. The more often a word is confirmed in a noisy corpora, the higher is its probability of being a graphically *correct* word form. 2. If these words can be altered in their morphology, words with a high frequency are *stems* or *lemmata*.

3 Acquisition of Lexical Knowledge

3.1 Frequency Lists

Since the introduction of Zipf's law [6] one of the most simplest, but nevertheless powerful methods of finding statistical regularities is to build a frequency list (see Table 1) or rank-frequency distribution. Due to the small size of the corpora and especially the large number of "noisy" words, which enlarge increadibly the number of single-occurence words (or hapax legomena), the conditions for a proof of Zipf's law are not adequate. As already pointed out in 2.2, correct and

rank	freq	word(s)	rank	freq	word(s)	rank	freq	word(s)
1	210	to	11	109	I, we	21	49	copy
2	209	the	12	105	annual	22	47	could
3	205	your	13	89	report	23	44	on
4	201	of	14	85	be	24	42	company
5	169	you	15	81	send	25	41	information, mailing
6	164	and	16	71	our	26	40	are, this
7	145	in	17	69	if	27	39	list, with
8	121	for	18	62	reports	28	38	thank
9	120	would	19	60	please, as, us	29	30	latest, address
10	111	a	20	54	is me	30	29	any

Table 1. Rank-Frequency List (first 30 ranks)

stem forms and here especially words with a domain specific meaning appear on the higher ranks. So frequency can be taken as a decisive characteristic to find the lemma for a number of correct or incorrect variants. The remaining problem is to find a way to subsume the variants under the more frequent stems and to measure the similarity between words in order to find the necessarily derivable and possible modalities of a stem. The following section shows a simple but efficient solution for this problem.

3.2 Levenshtein Distance

The most frequent problems caused by OCR (e.g. Merging, Splitting or Replacement of characters; incorrect word boundary recognition) can be captured using a method based on the Levenshtein distance that can also be used to determine lexical similarity [3]. Two words (with the same or different lengths) are compared with each other in a distance matrix, which measures the least effort of transforming one word into the other. Least effort means the lowest number of insertions, deletions, or replacements (as a combination of deletion and insertion). The effort is normalized to the length of the longest word to obtain a ratio-scaled value. Table 2 shows the unordered lexicon entries for the word form *rbport* with all similar words that were found in the corpus, having Levenshtein distance lower than 1.0. Against all expectations and as already proved in sec-

word	variants	distance	freq	variants	distance	freq	variants	distance	freq
rbport $freq = 1$	report	0.333	89	xport	0.666	1	importance	0.8	1
	reports	0.428	62	sports	0.714	1	portfolios	0.777	1
	roports	0.428	1	cort	0.666	1	opportunity	0.818	3
	ofreports	0.555	1	fjeport	0.714	1	north	0.833	1
	reporting	0.555	2	portfolio	0.777	1	opportunities	0.846	2
	reporting"	0.6	1	important	0.777	1			

Table 2. Unordered lexicon entry

tion 2.2 the number of correct forms and "deflected" forms is always higher than those of *typical* OCR-mistakes. In fact it must be asked, whether typical OCR mistakes exist at all due to the different types of reasons for these mistakes and the multitude of effects they may have.

3.3 The Core Lexicon

The lexicon that consists so far of all unordered lexicon entries (s. Tab. 2) has very low structure and consists of entries for all types and word forms having a certain similarity to them. No differentiation concerning lemmata and variants is done. To reduce the number of entries and to bring some order to the lexicon, the results of the frequency lists and Levensthein distances $d_{(s/v)}$ are combined as follows.

The algorithm processes successively through the frequency list, starting with the most frequent word and finishing with the last hapax legomenon. Each word that can be found in the frequency list is considered as the top of a new lexicon entry or lemma. Afterwards, the algorithm looks for the word forms in the lexicon, that are similar to this word, assigns them as variants in the new entry and recursively looks for all variants of the previously assigned variants. Each one of these variants can no longer be regarded as top of another entry and consequently is taken out of the frequency lists, that simultaneously shrinks more and more. The variants frequency is added to that of the lemma. The effects

of the algorithm depend a lot on an *a priori* specified threshold value for the Levenshtein distances. In our tests, good results are achieved with a value of 0.45 for direct similarity and 0.7 for indirect similarity, meaning the newly computed distance of variants of a variant to a given lemma.

The result of this process is a core lexicon that consists of a) high frequent synsemantica or function words having no variants, b) high frequent, domain specific autosemantica or content words and most of their occuring variants, c) middle and low frequency words and their variants, and d) one single entry for all the remaining hapax legomena having no similarity to one of the preceding words lower than 0.45, in order of their summarized frequencies. Hence, the number of entries in the core lexicon is at about one third of the total number of types. Table 3 shows some of the domain significant entries for the English corpus. As follows, many of the wrongly analyzed combinations of *your annual report* that lead to a rejection of the text, now can be transformed into their correct forms. This increases the number of documents that can be analyzed by the system considerably.

stem	variants	$d_{(s/v)}$	freq	stem	variants	$d_{(s/v)}$	freq	stem	variants	$d_{(s/v)}$	freq
your			205	annual			105	report			89
	yours	0.2	1		"annual	0.142	2		reports	0.142	62
	youor	0.2	1		iannual	0.142	1		reprt	0.166	2
	ofyour	0.333	2		annuai	0.333	4		repo	0.333	1
	z&your	0.333	1		annuad	0.333	1		rbport	0.333	1
	yours.	0.428	1		annua#	0.333	1		ofreports	0.333	1
					semiannual	0.4	1		reporting	0.333	2
					yourannual	0.4	1		reporting"	0.4	1
									roports	0.428	1
									fjeport	0.428	1
Σ	5		211		7		116		9		161

Table 3. Lexicon entries: *your, annual, report*

A comparision of the core lexicon with common frequency analyses [2] for correct texts shows that even with a very small text sample the resulting information for linguistically allowed alterations or lemmatizations of a lexical base form are achieved, at least in restricted domains. Additional information is achieved with the subsumption of linguistically incorrect variants. Thus, the core lexicon does not only bear the basic lexical knowledge that is needed for a "robust lemmatization" of a given text but furthermore enables the "cleaning" of documents from noisy sequences as can be seen in Fig. 2: 1. The words in normal print are identified as a lemma already existing. 2. The words in italic are identified as a variant of a lemma already existing and is lemmatized (italic print). 3. The word is neither a lemma nor a confirmed variant and is compared with the base entries (bold print). 4. The words (in parenthesis) are not recognized as similar to one of the core lexicon's entries.

Our collecon of buslness roports and acc0unts is an important and well-used resource for staff and students of our Business Sc oo and we feel that as many companies as possible should be represented. We shouid therefore be very grateful if ou could send .us a copy of all reports since 199-2 and re-add our name to your maiing list to recieve them in future.	our **collection** of *business report* and **accounts** is an important and (well-used) *resources* for staff and students of (our) business sc (oo) and we feel that as many *company* as possible should be **present** . we **should** therefore be very grateful if ou **would** send (.us) a copy of all *report* since **1992** and (re-add) our name to your **maiing** list to *receive* them in future .

Fig. 2. Robust Lemmatisation of an unknown Text

4 Conclusions

The paper reflects a quantitative approach to language independent lexical acquisition considering the sparseness and noisiness of certain text corpora. Word forms are aligned automatically to their lemmata with a set of very small, but domain specific texts. First results show that the core lexicon that is learned during the training process can be used to "clean" documents from noisy sequences. It builds the basis for the processing of new documents, dynamically enlarging with each new text. Meanwhile the learning process continues and converts into a rather symbolic approach with the determined lemmata as patterns. Each new variant is assigned to the corresponding lemma.

References

1. T. Bayer, U. Bohnacker, and I. Renz. Information extraction from paper documents. In H. Bunke and P.S.P. Wang, editors, *Handbook on Optical Character Recognition and Document Image Analysis*, pages 653–677. World Scientific Publishing Company, 1997.
2. W.N. Francis and H. Kučera. *Frequency Analysis of English Usage.* Houghton Mifflin, Boston, 1982.
3. J. Nerbonne, W. Heeringa, E. van den Hout, P. van der Kooi, S. Otten and W. van de Vis. Phonetic distance between dutch dialects. In Durieux, G., Daelemans, W., and Gillis, S., editors, *Proceedings of Computational Linguistics in the Netherlands*, pages 185–202, Antwerp, Centre for Dutch Language and Speech (UIA), 1996.
4. C.E. Shannon. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:623–656, 1948.
5. E. von Weizsäcker. Erstmaligkeit und Bestätigung als Komponenten der pragmatischen Information. In E. von Weizsäcker, editor, *Offene Systeme I*, pages 83–113. Klett, Stuttgart, 1974.
6. G.K. Zipf. *The Psycho-Biology of Language.* Houghton Mifflin, Boston, 1935.