Webaffix: Discovering Morphological Links on the WWW

Nabil Hathout and Ludovic Tanguy

ERSS / CNRS & Université de Toulouse Le Mirail - France
5, allées A. Machado, F-31058 Toulouse CEDEX 1
{hathout,tanguy}@univ-tlse2.fr

Abstract

This paper presents a new language-independent method for finding morphological links between newly appeared words (i.e. absent from reference word lists). Using the WWW as a corpus, the Webaffix tool detects the occurrences of new derived lexemes based on a given suffix, proposes a base lexeme following a standard scheme (such as noun-verb), and then performs a compatibility test on the word pairs produced, using the Web again, but as a source of cooccurrences. The resulting pairs of words are used to build generic morphological databases useful for a number of NLP tasks. We develop and comment an example use of Webaffix to find new noun/verb pairs in French.

1. Overview

We present a new language-independent method of finding morphological links between newly appeared words (i.e. absent from reference word lists).

Our reflection originates from the observation that only a few languages have easily available morphological databases. The only widely distributed one is CELEX (Baayen et al., 1995) for Dutch, English and German. In particular, no such database is available for romance languages, especially in terms of coverage and price. Besides, these databases need a costly updating process. What we present here is a way to extend such existing databases, or even to build them from scratch. In order to avoid an expensive (and error-prone) manual selection of lexical units, we propose a way to obtain them from the Web. As the results of this kind of search in a corpus as chaotic as the WWW are very noisy, we will therefore propose and evaluate a method for filtering candidates words, thus reducing the amount of human work needed.

The two main hypotheses behind our method are:

- Newly constructed words use a finite set of suffixes for each grammatical category (for languages with a concatenative morphological structure). Thus it is possible to get access to new words through suffixes, and their base lexeme can easily be guessed.
- A good clue for the identification of such constructed words is the presence of the base lexeme in the same text (or Web page in our case). This principle has been used on regular corpora (Xu and Croft, 1998).

As indicated by the second hypothesis, the use of a corpus is central in our approach. We will first have to discuss the consequences of having chosen the WWW as a corpus, as it is in several important ways very different from a regular corpus.

The example we will present in the course of this article is the search for French deverbal nominals, along with their base verb. Example pairs discovered by this method are: *(covoiturage/ covoiturer*[1]*), (ecz´ematisation/ecz´ematiser*[2]*), (pacsage/pacser*[3]*).* As can be seen, the word pairs found are those that do not appear in lexical databases because of their relative novelty, and technical or slang status.

The usefulness of such information is well known in computational linguistics. The most obvious use is for IR systems, where derivational information provided by a database such as CELEX can be used for query expansion (Jing and Tzoukerman, 1999; Jacquemin and Daille, 1998). But there are other fields of NLP concerned with such resources, such as syntactic analysis, where the argumental structures known for the verb can be inherited by the related noun, in pairs on sentences such as *l'**atterrissage** de l'avion*[4] and *l'avion **atterrit*** [5](Bourigault and Fabre, 2000).

In both cases, there is no need for semantic information on the link between the two lexemes, only the derivational status is needed. That is why Webaffix's result only consists of lists of such pairs.

2. The Web as a morphological and lexical resource

Using the WWW as a corpus is a current trend in the acquisition of lexical resources (Kilgarriff, 2001): its sheer size and the lexical creativity one can find on Web pages are a good enough counterpoint to its disorganization and lack of reliability (Grefenstette, 1999). Our approach here is to limit our research to lexical items for a given language, and to focus on getting reliable information, rather than as many new words as possible.

The main characteristics of the Web as a corpus that we deem relevant to our study are:

- **Search engines -** The only available way to access the Web on a large scale is to use generic search engines. These tools are of course limited to dealing with surface lexical forms, and only give access to a part of the Web. Another technical point which will be discussed

[1] Ride sharing.

[2] Medical words for "to develop an eczema".

[3] Slang words referring to the newly-adopted French PACS (administrative status for living with another person without being married).

[4] The landing of the plane.

[5] The plane lands.

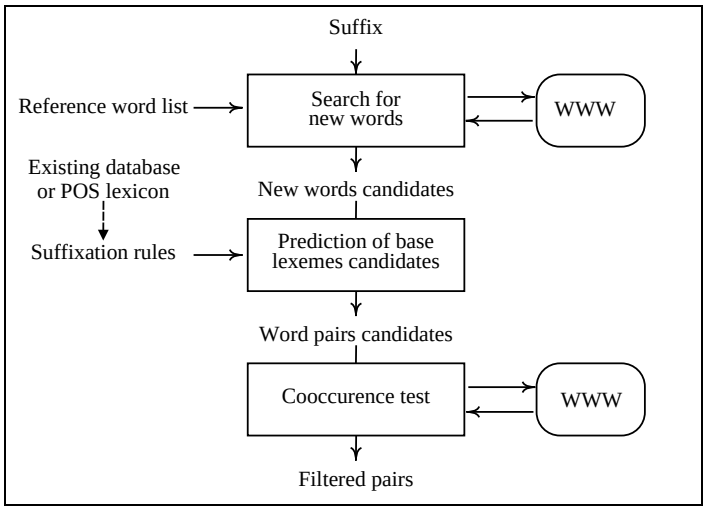

Figure 1: Overview of Webaffix's modules

below, is the lack of features in most search engines for looking for unknown words in specific languages.

- **Open-endedness -** The Web changes, and it cannot be considered as a finite corpus. The main consequences for our purpose is that we cannot get access to the complete list of lexemes occurring in the corpus, and that the frequency of these lexemes remains unknown[6]. These points prevent the use of statistical methods for the morphological matching of words, such as the ones used in (Yarowsky and Wicentowski, 2000).

- **Granularity -** The notion of document in the WWW is tightly linked to that of web page. This partition of the corpus is made obvious by the use of search engines, and does not match the classical notion of document in a traditional corpus. What we would call a text can be divided into several web pages for publishing purposes, or can be combined with several other unrelated texts to form a heterogeneous page. The cooccurrence test we use in Webaffix to filter word pairs will have to cope with this fuzzy notion.

- **Heterogeneity and lack of information -** The most annoying and frightening features of the WWW for a corpus linguist come from the "pot-pourri" status of the corpus. We have no control on the kind of texts that we are given access to, and we have also no information about the texts we process. It would be an obvious mistake to consider the WWW as a generic linguistic corpus. Even if Kilgarriff (Kilgarriff, 2001) sees it as the next step after the BNC in the history of corpus linguistics, we are still far from a similar description of its components (at least in terms of extra-linguistics information such as author's status, time period, intended audience, etc.)

[6]WWW search engines only give a rough approximation in terms of the number of Web pages where a given word appears, not the actual frequency of the word.

The consequences of using the Web in our study can be summarized in the following points:

- We must restrict ourselves to the study of lexical, information poor phenomena
- We cannot use nor rely on the frequency of tokens
- We need to filter the sheer output of search engines
- We must deal with a rather crude compatibility test
- As an advantage, though, our search on the WWW leads to rare and new words

3. The Webaffix tool

3.1. Overview of Webaffix

The method described here is divided into three steps. Only the first and last use the WWW as a resource, while the second one relies on existing morphological resources (used through bootstrapping) or, if such resources are not available, morphological rules. Webaffix is meant to be run iteratively, taking advantage of the accumulated data, and of the fact that the Web is an everchanging corpus, thus leading to different results through time.

The three steps are the following :

1. Search for new words ending with a given suffix. This step uses WWW search engines, and an optional reference list for filtering out already known words.
2. For each resulting word, production of a hypothetical base lexeme (*e.g.* a verb if the goal is to find nominals ending with a given suffix). The predicted base lexeme is also inflected if needed.
3. For each such pair, search for web pages where both the base and derived lexemes appear.

An overall picture of Webaffix's modules is presented in figure 1.

We will now describe the three modules.

3.2. First Module: looking for new words

As stated before, the search for new words starts with a suffix, or list of suffixes to be found at the end of words. For romance languages such as French, which have a concatenative morphology, it is legitimate to assert that lexical creation uses a few derivational schemes, and so a few possible suffixes. Our experiment deals with the following French nominal suffixes, widely used for derivation from verbs: *-ade, -age, -ance, -ement, -erie,-tion.*

As we focus on newly appeared words, we need a reference list of existing words. We use an electronic lexicon obtained from the "Trésor de la Langue Française" (or TLF hereafter) that contains 514,871 inflected forms. Many such word lists can easily be obtained for a wide range of languages, as we do not need any additional information, such as part-of-speech tags or other lexical descriptions.

The first problem arises when we want to define WWW queries that look for words ending with a given substring. The wildcard (or $*$) feature is not available for all search engines, and even for the ones providing it, it has heavy restrictions.

For common uses of WWW search engines, the $*$ is used as a pseudo-stemming feature, allowing the user to use incomplete words in his/her query, for example asking for "*avoid*$*$" instead of typing (*avoid OR avoids OR avoiding OR avoidable...*). Our problem is different, as we need a wild card at the beginning of our query tokens.

No search engine provides such a computationally costly feature. The more tolerant one is AltaVista™[7], as it requires only three standard characters before allowing a $*$. The only other engine featuring this is Northern Light™[8], but it requires four characters instead of three. We chose to use AltaVista, but it means that a single query (*i.e.* for a single suffix) has to be divided into thousands of sub-queries. Still using our reference word list, we calculated the plausible sequences of three letters appearing in the beginning of French words, leading to 3,117 such combinations (including accented characters). For a given suffix such as *-tion*, this involves the creation of 3,117 queries, ranging from "aal$*$tion" to "zyt$*$tion". It would have been more than 13,000 if we had to produce 4-characters long subqueries.

Additionally, we restrict these subqueries by excluding known words matching the scheme (found in our word list), and by restricting the search to French pages. Both restrictions are performed by the search engine.

For practical reasons, in order to limit the computation (and network load) of this first stage, we only process the first 20 pages proposed by the search engine. For the six suffixes, this still led to the analysis of more than 120,000 web pages.

Each such page has to be downloaded and parsed, in order to find the matching word. This step also includes a series of filtering processes, as a large quantity of noise is still provided by the search engine, despite the restrictions mentioned.

The main sources of noise are:

- The web page is a dead link (around 6% of pages)
- The web page has changed since its indexing by the search engine's crawler, or the word is absent from the main text of the page (around 26 % of pages)
- The occurrences found are not in the target language (around 17% of occurrences) This is due to either:

  - a misclassification of the page's language by the crawler

  - a multilingual page, with the matching words appearing in a segment of text written in another language

  - occurrences belonging to a closely related language, such as Occitan or old French, for which no language detection procedure exists for generic crawlers.

- The occurrences are not words, but parts of computer code, URL, mail addresses, etc. (around 12% of occurrences)
- The occurrences are misspelled (around 35% of occurrences).

We had to deal individually with every source of noise. While we could do nothing for the first two, the three others are partially filtered out with shallow analyses. We implemented a rough language detection routine, that takes into account a small window (50 words) around the target occurrence, relying on the presence of foreign stop words. Specific case and punctuation marks are used to detect codes and URLs problems. The spelling errors are quite difficult to take care of, as we cannot use a spelling checker, for our target words are most of the time absent from any reference lexicon. We only applied lightweight detection routines based on (badly) accented letters and sequences of identical letters.

A last source of errors is inherent to the method we use. For a given suffix, we are of course expecting words belonging to a given part of speech category (nouns in the case we describe here). In most cases, there are valid words that match the query but belong to another class, such as adverbs for the French suffix *-ement*. We tried using a POS tagger on the analysed web pages, but had a problem similar to the spelling errors: generic text analysis tools perform very badly on unknown words and, in our case, lead to tagging them as nouns. This problem will be partly solved by the other two modules.

Out of the 120,000 web pages analysed, the first Webaffix module produced a total list of 13,500 words for all of the 6 suffixes. Of course, some suffixes are more productive than others, and lead to different kinds of errors. Table 1 presents an overview of the results for each suffix, along with an evaluation of the remaining errors, based on the manual evaluation of a random sample of 100 words for each suffix.

As shown in the figures of table 1, the most important remaining source of noise is due to spelling errors, which cannot easily be dealt with in our case. The foreign context detection still needs to be improved, but most of the time the errors are due to very short segments of foreign text

[7] `http://www.altavista.com`
[8] `http://www.northernlight.com`

Suffix	*-ade*	*-age*	*-ance*	*-ement*	*-ence*	*-erie*	*-tion*	Total
Web pages (#)	11,618	22,599	15,132	20,261	9,945	12,951	27,664	120,170
Occurrences(#)	2,531	10,970	4,395	10,350	3,828	3,193	12,162	47,429
Word forms(#)	813	2,189	1,097	3,791	999	995	3,564	13,448
Correct word(%)	29	66	40	17	34	59	53	45
Wrong category(%)	14	1	5	23	2	10	1	8
Foreign word(%)	36	7	14	6	11	11	18	12
Spelling error(%)	16	23	36	51	50	16	27	33
Code, URL, etc. (%)	5	3	5	3	3	4	1	3

Table 1: Results for the first module : Amount of data processed and evaluation of filtering

(such as translation of terms, or reference to the original title of a film/book, etc.) The main problem for French is the relatively high frequency of texts in Old French, whose surface features make it quite difficult to differentiate from modern French. As can also be seen, there are variations between the kind of errors across the different suffixes. This is mostly due to the similarity of words across languages (*-ade* is common in Spanish, and *-tion* in English). The number and percentage of correct words are correlated to the productivity of suffixes.

However, these figures are quite satisfactory, as most of the filtering will be performed by the third module (looking for cooccurrences). The amount of network traffic and text analysis required for the first module is quite high. The results presented here were obtained during the first week of February 2002, with an overall computation time of 150 hours. This relatively high amount is essentially due to network traffic, as the filtering components are very light.

3.3. Second Module: base lexemes prediction

The list of words obtained from the first module is passed through to the second one, without any kind of human intervention. Thus, the prediction module has to deal with candidate nouns. Some of them, it has been noted, are not nouns, most of them are misspelled, and some of them are foreign words.

The way this module performs its prediction is through a learning technique, taking advantage of an existing French database of noun-verbs pairs, named Verbaction[9] This method consists, for a given word, in comparing it to existing word forms in order to select the ones ending with the longest identical substring. For example, the new word *d´esaffixation* will be matched against the *fixation/fixer* pair in the Verbaction database (although there is no morphological link whatsoever between *fixation* and *d´esaffixation*), and lead to the candidate base verb *d´esaffixer*. The analogy principle is followed up to the inflection of the base lexeme, thus dealing efficiently with allomorphy; we do not use any rule-based inflection algorithm, but use instead a lexicon that contains fully inflected forms.

In the case of a new database, suffixation schemes could be learned from a morpho-syntactic lexicon only. This learning technique has been used by (Dal et al., 1999) and (Gaussier, 1999).

The result of this first step is a list of inflected verb forms, such as *fixer, fixe, fix´e, fixez, fixons,* etc.

The last step is to filter the possibly ambiguous verb forms. If we blindly follow the previous process, we can obtain strings that can be used as nouns, or adjectives. An example is the new word *afficherie* (advertising poster workshop), which leads to the base verb *afficher* (to display). The inflection of *afficher* (first person present indicative) gives *affiche* (I display), which is homomorphic to the noun *affiche* (advertising poster). When looking for cooccurrences, we will get wrong results, as we will have no means to disambiguate the occurrences of *affiche*. So, every inflected form obtained through this method, and which appears in our reference lexicon with a POS other than verb is filtered out. Of course, we cannot avoid whole families of new words, and sometime get wrong results because of the homomorphy with an unknown noun or adjective.

The last step in this module is quite trivial, as it consists in building a complete Boolean query that will be sent to the WWW search engine by the last module. A sample query is indicated in figure 2. It is important to note that these queries can be quite long, and this led us to use AltaVista™, as it allows queries to be up to 800 characters long (please note that we have no commercial interest in this company!)

```
(pacsage OR pacsages) AND (pacsa OR pacsai
OR pacsaient OR pacsais OR pacsait OR pac-
sant OR pacsas OR pacsasse OR pacsassent
OR pacsasses OR pacsassiez OR pacsassions
OR pacse OR pacsent OR pacser OR pacsera
OR pacserai OR pacseraient OR pacserais
OR pacserait OR pacseras OR pacserez OR
pacseriez OR pacserions OR pacserons OR
pacseront OR pacses OR pacsez OR pacsiez OR
pacsions OR pacsâmes OR pacsât OR pacsâtes
OR pacsèrent OR pacsé OR pacsée OR pacsées
OR pacsés)
```

Figure 2: Sample query built by the second module for *pacsage*

3.4. Third Module: compatibility test with cooccurrences

The third modules also uses the Web as a corpus, as its point is to look for cooccurrences of both the base lexeme

[9]This database contains around 7,000 noun-verb pairs for French, and has been manually validated. The words described in this database were selected from the TLF dictionary.

Suffix	*-ade*	*-age*	*-ance*	*-ement*	*-ence*	*-erie*	*-tion*	Total
Candidates	813	2,189	1,097	3,791	999	995	3,564	13,448
Selected	55	450	154	385	81	85	611	1,821
Selected(%)	6.77	20.56	14.04	10.16	8.11	8.54	17.14	13.54

Table 2: Results for the third module

and its derived form. As will be seen from the resulting figures, it is during the run of this module that most of the filtering is processed. The hypothesis beyond this test is that newly constructed lexical forms appear in the vicinity of the base lexeme for two reasons:

- In the case of a raw creation, the author takes care of insuring the reader's comprehension of his/her neologism by explicitly indicating the base form. This often leads to some kinds of explicit definitions, *e.g.*:

  *Une ligne **quotée** est une ligne avec un signe de **quotage**.*

  Most of the time the base lexeme appears after the construction, without explicit link, but with a cooccurrence nevertheless.

- In the case of technical jargon, the amount of repetition is so high that it always leads to using both the base and the constructed lexeme in the same document, with alternating occurrences, *e.g.*:

  *Si l'Etat **cofinance** l'achat (...) Dans le cas d'un **cofinancement** par l'Etat...*

- In the case of generic vocabulary, both forms alternate, as a way to avoid unstylish repetitions.

The compatibility test we use is very simple. We require a word pair to be present in at least one document from the WWW in order to be selected. In this process, we apply the same filtering rule as those described for the first module, in order to filter out non-text segments or foreign languages. The queries used are the ones built by the second module.

An interesting source of noise appeared at this stage and had to be dealt with. There exists on the WWW a large number of Web pages which contain (sometimes exclusively) lists of unrelated words. The purposes and natures of such documents are many, ranging from lexical resources developed by computational linguists and distributed on the WWW to password-cracking word lists used by hackers, and of course pornographic web sites that lure the search engines' crawlers to index them with every possible keyword. In all of these cases, obviously, the notion of cooccurrence between the two lexemes is not a clue as to the validity of the morphological link. We can easily and automatically detect these web pages, as most of the time the words appear in alphabetical order.

Out of the 13,448 candidate nouns we obtained from the first module, only 1,821 couples were selected with this method. The differences between the different suffixes are described in table 2.

These figures show that the most effective filtering is achieved through the third module, which leaves only 13% of the candidates to the manual checking in the end.

It should be noted here that our goal is to reduce the amount of time dedicated to manual selection of word pairs, and as such we seek precision instead of recall. We may leave aside valid candidate pairs with such a drastic method, but cannot estimate the loss of such information from the beginning of the selection process. The WWW search engines themselves do miss interesting documents in the first place. As stated above, Webaffix can be run iteratively, taking into its reference database the results of previous runs, and at each time using a slightly different corpus.

The final evaluation of this module will be described in the next section.

4. Evaluation

As an evaluation for the filtering processes of Webaffix's first module has already been presented, what we present here is the quality of the whole set of processes. The 1,821 resulting word pairs have been manually evaluated, leading to the classification of each one in the following categories:

- **Wrong POS.** The candidate word does not belong to the target part of speech category. This source of error has been described along with the first module. The number of incorrect results is much lower after the third step than it was after the first one, though.
- **Foreign language.** This problem is still present, but with a significant difference. Most of the incorrect pairs of words are in fact well-formed derivations, but in the wrong language. The hypothesis of cooccurrence is valid, even if the whole documents are in old French. We thus get linguistically interesting information, but irrelevant for our study, such as the *fascherie/fascher* pair, with correspond to the modern *fâcherie/fâcher* (quarrel/to anger)).
- **Spelling errors.** Again, the relative proportion of spelling errors has been lowered, but is the most important source of errors.
- **Wrong semantic link.** This last kind of error appears only after the third step. It means that the resulting word pairs do not have the semantics we are looking for, *i.e.* the noun is not a nominal derived from the verb. Such cases are heavily dependent on the suffix, for example the *-erie* suffix is often used to derive the name of the place where the action is performed, instead of the action itself.

The detailed error rates for each suffix are presented in table 3.

The results are more than encouraging, especially for the more productive affixes such as *-age* and *-tion*. Although the number of resulting pairs may be found low

Suffix	*-ade*	*-age*	*-ance*	*-ement*	*-ence*	*-erie*	*-tion*	Total
Pairs	55	450	154	385	81	85	611	1821
Correct (%)	**24**	**79**	**20**	**32**	**11**	**16**	**68**	**52.76**
Wrong POS (%)	7	1	1	1	1	4	1	1.32
Foreign language (%)	33	4	18	12	1	26	3	8.31
Spelling errors (%)	16	16	42	43	65	12	27	29.59
Wrong semantic link (%)	20	0	19	12	21	42	1	7.98

Table 3: Evaluation of resulting pairs

compared to the number of candidates, one has to remember that our goal is to get as reliable an information as possible. This is all the more important, given the lack of reliability and control we have on the web as a corpus.

5. Conclusion

The method presented here is a very general one. It can be used for any language that follows the same morphological principles, and needs only very limited resources. The resulting morphological links can be used in a number of NLP applications, among which information retrieval.

A further development will consist in improving the error detection routines of the first modules, especially for dealing with closely-related languages (dialects or old variants).

Another important point will be to deal with the heterogeneity of the harvested word pairs. As shown in the examples, we get both slang words and technical jargon, and of course these two categories should be distinguished. This can only be done through a more thorough analysis of web pages as documents, perhaps using extra-linguistic information specific to web pages, such as the overall hypertextual structure of the page, amount of pictures and coloring, etc. Some work is in progress toward such a characterization (Beaudouin et al., 2001).

Another interesting, more linguistically-oriented perspective is the investigation of the use of different suffixes in different situations,*e.g.* some technical or science fields seem to prefer a given suffix, as part of their sociolectal habits. Both of these directions lead to a better characterization of the Web as a corpus, but with the help of linguistics itself.

6. References

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX lexical database (release 2). CD-ROM. Linguistic Data Consortium, University of Pennsylvania, Pennsylvania, USA.

Valérie Beaudouin, Serge Fleury, Benoˆıt Habert, Gabriel Illiouz, Christian Licoppe, and Marie Pasquier. 2001. Typweb : Décrire la toile pour mieux comprendre les parcours. In *CIUST'01 : Colloque International sur les Usages et les Services de Télécommunications*, Paris.

Didier Bourigault and Cécile Fabre. 2000. Approche linguistique pour l'analyse linguistique de corpus. *Cahiers de Grammaire*, 25:131–151.

Georgette Dal, Nabil Hathout, and Fiammetta Namer. 1999. Construire un lexique dérivationnel : théorie et réalisation. In Pascal Amsili, editor, *Actes de la VI^{e} conférence sur le Traitement Automatique des Langues Naturelle (TALN'99)*, pages 115–124, Cargèse, Corse, jul. ATALA.

Eric Gaussier. 1999. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the Workshop on Unsupervised Methods in Natural Language Processing, Association for Computational Linguistics, ACL'99*, Univ. of Maryland.

Gregory Grefenstette. 1999. The www as a resource for example-based mt tasks. In *Proceedings of the ASLIB 'Translating and the Computer' Conference*, London. Invited Talk.

Christian Jacquemin and Béatrice Daille. 1998. Lexical database and information access: A fruitful association. In *Proceedings, First International Conference on Language Resources and Evaluation (LREC'98)*, pages 669–673, Granada.

Hongyan Jing and Evelyne Tzoukerman. 1999. Information retrieval based on context distance and morphology. In *Proceedings of 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 90–96, Berkeley, CA. ACM.

Adam Kilgarriff. 2001. Web as corpus. In *Corpus Linguistics 2001*, Lancaster.

Jinxi Xu and W. Bruce Croft. 1998. Corpus-based stemming using co-occurrence of word variants. *ACM Transaction on Information Systems*, 16(1):61–81.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceeding ACL 2000*, pages 207–216, Hong-Kong.