RC 22029 (W0104-009) April 11, 2001
Computer Science

IBM Research Report

Derivational Morphology to the Rescue: How it can help resolve unfound words in MT

Claudia Gdaniec, Esmeralda Manandise, Michael C. Mc Cord
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Derivational Morphology to the Rescue: How It Can Help Resolve Unfound Words in MT

Claudia Gdaniec, Esmé Manandise, Michael C. McCord

IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598
USA
{cgdaniec, esme, mcmcord}@us.ibm.com

Abstract

Machine Translation (MT) systems that process unrestricted text should be able to deal with words that are not found in the MT lexicon. Without some kind of recognition, the parse may be incomplete, there is no transfer for the unfound word, and tests for transfers for surrounding words will often fail, resulting in poor translation. Interestingly, not much has been published on unfound-word guessing in the context of MT although such work has been going on for other applications. In our work on the IBM MT system, we implemented a far-reaching strategy for recognizing unfound words based on rules of word formation and for generating transfers. What distinguishes our approach from others is the use of semantic and syntactic features for both analysis and transfer, a scoring system to assign levels of confidence to possible word structures, and the creation of transfers in the transformation component. We also successfully applied rules of derivational morphological analysis to non-derived unfound words.

Keywords

Unfound words, derivational morphology, computational morphology, transfer system, scoring mechanism

1 Problem

Natural language processing (NLP) systems whose parsers rely on the morphological and syntactic information associated with words, such as part of speech (POS), gender, number, case etc., run into problems when words are not found in the lexicon. Generally, such unfound words contribute to the bad parsing of input text. In MT systems, these words are often resolved as nouns, sometimes flagged as unfound, and never translated. Apart from possible bad parsing, the semantic information for determining the correct transfer of surrounding words is also missing.

2 Motivation and Purpose

Updating the lexicon would be the most direct way to remedy this problem. However, this is a time-consuming and costly enterprise. And since productive word-formation rules allow the creation of new words almost on a daily basis, no lexicon is ever complete.

Our linguistic intuition tells us that knowledge of word structures and the regularity of word formation can be formalized into a procedure to deal with unfound words not only for an improved analysis of the input but also to generate a more meaningful MT output.

We implemented a language-independent strategy for responding to unfound words in arbitrary unrestricted texts through the morphological analyzer (MA) of the MT system LMT developed at IBM, whose main application is Web-page translation in several languages. Another crucial part of this work was the generation of transfers in our transformations. For the work described in this paper, French and German[1] were chosen as source languages and English as the target language.

3 Related work

Researchers in NLP have developed MAs that incorporate derivational morphology as a strategy to improve application performance. See for example: general purpose analyzers (Byrd et al. 1986), analyzers for improved tagging and parsing (Daciuk, forthcoming), and for enhanced information retrieval and conceptual indexing (Klavans, Jacquemin, & Tzoukerman, 1997; Woods, 2000). Morphological analysis is also crucial for improved spell checkers and text-to-speech applications (see Sproat, 1992). The extent of the use of derivational morphology seems to depend on the application[2]. MT is the only application that needs not only to analyze the input but also create transfers.

In MT, publications on how MT systems handle unfound word-formation are rare (Hutchins & Somers, 1992; McCord & Wolff, 1988). One reason may be the fact that MT always needs acceptable transfers, which are difficult to generate for unfound words.[3]

[1] We discuss only words derived through affixes or conversion. Compounds are not mentioned here.

[2] For general descriptions, see Klavans & Tzoukerman (1992) and Sproat (1992).

[3] Tucker (1987, p. 26) argues that derivational morphology for MT is a theoretically uninteresting issue. Arnold et al. (1994, p. 99), on the other hand, suggest that the transfer part is "currently more of a research goal than a practical possibility."

4 Brief Description of the LMT MA

The LMT MA is a non-deterministic analyzer written in C. It consists of three steps:

1. affix stripping and base spelling adjustment or stem change
2. lexical lookup
3. affix operations

In step 1, an input word is subjected to language-specific morphological operations which handle inflectional and derivational morphology. The output is a list of *word structures*, consisting of possible base words with affix lists. Regardless of the type of affix (the program handles suffixes, prefixes, infixes, circumfixes, umlauts, as well as combining forms for German), the analyzer attempts to match a substring taken from the beginning or the end[4] of the input word against context-sensitive rules. Several different substrings can be isolated in a word in an iterative process.

In step 2, the hypothesized base words are looked up in the lexicon[5]. If found, they are returned to be processed in the third step.

In step 3, rules apply to the affixes to decide whether the combination of a base word and the affix(es) yields a valid word of the language. For inflectional affixes, the operations assign morphosyntactic features. For derivational affixes, they can assign a part of speech, morphosyntactic features, semantic features, and syntactic arguments. The rules create a bracketed word structure, which is passed on for later transfer.

5 Analysis and Transfer of Unfound Words

5.1 Principles and Definitions

Our approach assumes the principle of compositionality for both the analysis and transfer of unfound words. We distinguish between derived words whose base is in the lexicon (derived words with lexical base), those whose base is not (derived words with unlisted base), and those whose base is unfound but shared with other derived words in the lexicon (words derived from related words).

5.2 Goals

The analysis must

- list among the base words at least one base from which the input word can be derived
- create complex, bracketed word structures
- set a part-of-speech with morphosyntactic features, and, where possible, semantic and syntactic features

[4] One exception is for German nouns where an umlaut is the only indication of plural. In this case, the substring is empty and the word structure consists of a changed stem and a meta-affix called "umlaut".

[5] Our lexicons are not full-form lexicons. They do not contain inflected forms of words but only one entry per word, with further stems for some verbs.

- retrieve a pointer to transfers of the words related to the unfound word

The transfer must
- create a target-appropriate string or subtree
- integrate the transfer and its modifiers correctly into the target tree

5.3. Derived Words with Lexical Base

5.3.1 Analysis

Let us look at the German noun *Darstellbarkeiten* (*representability/portrayability*_ plural) to illustrate how the analysis proceeds. Step 1 returns affixes (prefix `dar`, derivational `bar` and `keit`, and inflectional `en`) and four possible base words to be looked up in the lexicon in step 2 (*stell, stellbar, darstellbar, stellbarkeit*). Step 3 examines the compatibility of the affix(es) with the base word(s) from the lexicon.
The word would be analyzed with word structure:

```
(1) ([darstellbar + keit]
    feminine_noun plural
    noun_object (penalty 1)
    "property")
```

In (2) below, we show parts of a simplified transcribed rule for the suffix `keit`, where language-specific properties and conditions are tested and features set.

(2) **`case s_keit:`**

```
if a word has been found
 without derivational affixes,
    then fail;
 else if this affix co-occurs
    with an umlaut,
    then fail;
 else if the POS is not
 adjective,
    then fail;
    newPOS = noun;
    gramm_gender = feminine;
 inflection_pattern = default;
 newSemantictype = "property";
 newWordstructure =
 wordstructure + keit;
 complements = noun_object;

 if word is dummyadjective,
    penalty = 10;
 else
    penalty = oldpenalty+1;

 if inflectional affix in
 affix list,
    do not set morphosyntactic
    features here;
 else if no other affixes in
 affix list,
    morphosyntax = singular
    nominative, genitive,
    dative, accusative;
```

Our system uses **penalties** as a measure to indicate confidence in the probability of appropriate analysis. For the often-cited English example *scarcity*, for instance, they would position the analysis [scarce+ity] higher on the preference list than the analysis [scar+city].

The following shows how penalties are assigned. The German noun *Darstellbarkeiten* is analyzed not only with the word structure shown in (1) above but also with a second word structure (3):

```
(3) ([[[dar+stell] +bar] +keit]
      feminine_noun plural
      noun_object (penalty 2)
      "property")
```

Penalty points are added incrementally with certain affixes. The second word structure is returned with two penalty points. The parse picks the first structure (1) because of the lower penalty score.

5.3.2 Transfer

The transfer of derived words with lexical base relies on
- Word structure transparency (bracketed structure from the morphological analysis is accessible to transformations)
- Pattern matching transformations on word structure[6]
- Base transfer is returned by the transfer shell
- String tests and/or tests on semantic features on target words

Let consider the French adverb *affirmativement*. Assume the adverb is not in the lexicon, but the verb *affirmer* is. Transformations are applied on the output from the syntactic analysis (4) to create (5) and (6) after the transfer shell replaces the French verb *affirmer* with the English *assert*.

```
(4)([[affirmer + tif] + ment]
      adverb "manner")

(5) Transformation Derived.3 on node #,
"assert", produces tree

|_adv assert ([POS adv] [Sslot
          vadv] [SSem "manner"])

(6) Transformation Manner_adv.1 on node
#, "asserting", produces tree...


|_____vadv   in ([POS prep])
 |    __ndet   a  ([POS det]
[Number sg]
                  [Person pers3])
 |   |__nadj assert ([VInfl ving]
 |   |                [POS verb])
 |   |
```

[6] See McCord & Bernth (1998) for a discussion of the LMT transformational system.

```
 |__|__objprep manner([POS noun]
[Number sg]
                    [Person pers3])
```

The final English output for *affirmativement* is *in an asserting manner.*

By creating the transfer structure in the transformations, which have access to syntactic and semantic features as well as to the transfer string itself, we are flexible with respect to the form of the new transfer. We can create one new word or a whole new subtree. *A parkable car* may be acceptable output, but *turnoffable* (*lights*) is less so. Transformations will instead create a relative clause for the adjective: (*lights*) *that can be turned off.* If the part of speech is changed, the POS and positions of its modifiers can also be changed in subsequent transformations. Nouns derived from verbs often sound more idiomatic when translated as an English gerund or ing-form. Transformations delete the definite article, change adjectives into adverbs, French prepositional *de* phrases into subjects or direct objects. The latter is based on subject and object preferences in the lexicon. For instance, transformations will transform the transfer "the reading of the child" to *the child's reading,* as opposed to "the reading of the book", which will be transformed as *reading the book.*

5.4 Derived Words with Unlisted Base

The analysis for derived words with unlisted bases is the same as that for derived words with lexical bases. However, since the success of an analysis depends on a successful match in the lexicon, step 1 returns a dummy base in some or all cases of derivational affixes, which is found in the lexicon (step 2). Which POS dummy is returned depends on the affix. In step 3, there are restrictions on dummy analyses as well as heavy penalties to make sure that they do not win over real derivations. The goal of making up a dummy word is to improve the parse when otherwise the word's part of speech would be determined incorrectly. The target generation consists in replacing the dummy target with the original input word.

Assume the misspelled German adjective *untenehmerische.* Step 1 returns, among other word structures, the derivational suffix `isch` and the following bases[7]:

(7) a. untenehmerische:
b. untenehmer: isch, e
c. dummynoun: isch, e

Step 2 does not find *untenehmer*, but finds *dummynoun.* After step 3 applies affix operations, an adjective is returned with the following word structure:

(8)
```
([dummynoun+isch]
adjective  (penalty 10)
"property")
```

[7] The words to the left of the colon are the hypothesized base words. The strings to the right of the colon are the affixes that are identified.

Adjective-specific transformations can now be applied on the output from the syntactic analysis. Instead of generating a target word from a base, a simple transformation retrieves the original input word and outputs it without any inflection. For the effect on parsing, compare (9a) against (9b) below.

(9) Input: Der sehr untenehmerische Vater hat ihm das Malen beigebracht.

(9a) is the output without dummy analysis and (9b) with the dummy analysis generated through derivational morphology.

(9a) This one very much, untenehmerische father taught him painting.
(9b) The very untenehmerische father taught him to paint.

5.5 Words Derived from Related Words

Based on the rightmost affix, other **related** words, which share a common root with the input word, are hypothesized as base words[8]. Assume the French adjective *admiratif* as input word. Word formation rules will hypothesize first the adverb *admirativement*, then the verb *admirer*, then the noun *admirateur*, then the noun *admiration* etc... The reason for starting with the adverb in this case is the closer distance between those two. This closeness is often reflected in the target. The strategy is being expanded to inflectional morphology, especially to those verbal suffixes which are nearly unambiguous. Assume that the French adverb *admirativement* is in the lexicon, but no other word with the root *admir-*. Our word-formation rules and procedures allow us to handle (10) below, where all the words in italics are unfound.

(10) Input: *Radmirons* l' *admirable admirateur admiratif* et l'*admirable admiratrice admirative.*
Old result:
Radmirons admirable admirateur admiratif and admirable admiratrice admirative.
New result:
Radmirons the admirable admiring admirer and the admirable admiring admirer.

6 Results

The strategy of deriving unfound words from base forms (including part-of-speech-specific dummy words) in the lexicon has good results but also some drawbacks.

6.1 Advantages

[8] According to a reviewer, Larry Piano and John Carroll developed a similar strategy in 1995/96 for guessing part-of-speech assignments for unknown words in an HMM part-of-speech tagger. For English, using a tagset of 160 tags, they got an accuracy of around 85% on unknown words as opposed to around 40% using a naive method (Personal communication, J. Carroll). This work was never published. It is available in a project-internal tech report (Piano 1996)

- Much larger coverage of unfound words if the lexicon is small
- Better parse
- In many cases acceptable transfers

(11) and (12) illustrate the above.

(11) Input: Les *admiratrices* de Sting devinrent folles.
Old result:
Admiratrices of Sting became crazy women.
New result:
The admirers of Sting became crazy.

(12) Input: Les *admirables* compositions musicales des Beatles appartiennent à la musique classique.
Old result:
Admirables musical compositions some Beatles belong to the classical music.
New result:
The admirable musical compositions of the Beatles belong to the classical music.

6.2 Drawbacks

- Less efficiency of processing time and space
- Sometimes transfers may not be easily understandable, e.g. (13) *jointly ownedness*.

7 Future work

For LMT, correct slot assignment is crucial for both syntactic analysis and transfer decisions. At the moment, we can transfer (part of) the argument structure of the base word to the derived word, which gives us improved translation by matching on co-occurrence tests. But a derived word may have different slots from those of the word it is derived from. For adjectives and adverbs, this is usually not an issue. It is sometimes a problem for nouns and adjectives derived from verbs as well as for those POS changes where the governance structure changes through derivation. These issues will have to be addressed in more detail in future work.

8 Conclusion

We have described how we use language-independent derivational morphology to improve MT output by analyzing unfound words, assigning morphosyntactic, syntactic and semantic features to them and then generating target words. The mechanism for generating transfers is flexible because it is performed in the transformations, with access to semantic features as well as words themselves. This allows us to go beyond word-by-word translation and to generate whole subtrees in the target. The quantitative impact of this work depends on the coverage of the specific lexicons. But even with an extensive lexicon, derivational morphology improves translation because every resolved part of speech, morpho-syntactic and semantic feature results in a better translation of a whole sentence, so that even with just 1% of the words unfound, 10% of the sentences in a document may get a better translation. We have also indicated how we use the mechanics of derivational morphology to recognize non-derived unfound words.

References

Arnold, D., Balkan, L., Meijers, S., Humphreys, R.L., & Sadler, L. (1994). Machine Translation. An Introduction. London: Blackwell.

Byrd, R.J., Klavans, J.L., Aronoff, M., & Anshen, F. (1986). Computer Methods for Morphological Analysis. In Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics (pp. 120-127).

Daciuk, J. (forthcoming). Treatment of Unknown Words. In Proceedings, WIA 99, Workshop on Implementing Automata. Berlin: Springer.

Hutchins, W. J. & Somers, H.L. (1992). An Introduction to Machine Translation. London: Academic Press

Klavans, J.L. & Tzoukermann, E. (1992). In S. Shapiro (Ed.), The Encyclopedia of Artificial Intelligence. 2nd Edition. New York: John Wiley & Sons (pp. 963--972).

Klavans, J.L. Jacquemin, C. & Tzoukermann, E. (1997). A Natural Language Approach to Multi-Word Term Conflation. In Proceedings, DELOS Workshop on Cross-Language Information Retrieval, ETHZ, Zurich.

McCord, M.C. & Bernth, A. (1998). The LMT Transformational System. In D. Farwell, L.Gerber, & E.Hovy (Eds.), Machine Translation and the Information Soup. Proceedings of the 3rd AMTA conference. Berlin: Springer (pp. 344--354).

McCord, M. & Wolff, S. (1988). The Lexicon and Morphology for LMT, IBM Research Division Research Report, RC 13403.

Piano, L. (1996). Adaptation of Acquilex tagger to unknown words – release 2'. Integrated language Database project work package CCL 2.1. Computer Laboratory. University of Cambridge.

Sproat, R.W. (1992). Morphology and Computation. Cambridge: MIT Press.

Tucker, A.B. (1987). Current strategies in MT research and development. In S. Nirenburg (Ed.), Machine Translation. Cambridge: Cambridge University Press (pp.22--41).

Woods, W.A. (2000). Aggressive Morphology for Robust lexical Coverage. In Proceedings of the Sixth Applied Natural Language Processing Conference (pp. 218-223).