MORPHOLOGICAL ANALYSER AND GENERATOR FOR BILINGUAL E-DICTIONARIES

Presented by

JISHA P JAYAN

Morphological Analysis

Morphological analysis is the segmentation of words into their component morphemes and the assignment of grammatical morphemes to grammatical categories and the assignment of the lexical morpheme to a particular lexeme or lemma.

There are different methods for the morphological analysis of natural language processing:

- ✓ **Brute Force Method**
- ✓ **Root Driven Method**
- ✓ **Affix Stripping Method**
- ✓ **Suffix stripping method**

The general format of the morphological analyzer is

Word → stem/root + suffixes

Reasons for morphological analysis

- **Identify newly encountered words.**
- **Extract roots for comparison of content.**
- **Determine parts of speech.**

Suffix Stripping Method

- Captures the creativity found in the inflectional system and analyze it.
- Its very economical.
- Analyser can analyze the inflected form of the word into suffixes and stem even though it is not present in the dictionary.

Makes use of

- Stem Dictionary
- Suffix Dictionary
- Sandhi Rules or Morphophonemic rules
- Morphotactics Rules

Bilingual Dictionary

- **Bilingual dictionaries are vital resources in many areas of natural language processing.**
- **In any machine translation system, the dictionaries are of critical importance, from two distinct aspects, their content and their organization.**
- **The content of the dictionaries must be adequate in both quantity and quality: that is, the vocabulary coverage must be extensive and appropriately selected and the translation equivalents carefully chosen if target language output is to be satisfactory or indeed even possible.**
- **The size and quality of dictionary limits the scope and coverage of a system, and the quality of translation that can be expected.**

Bilingual Dictionary

- **Indispensable working tools for translators and translation trainees.**
- **The dictionary entries are based on dictionary entries for lexical stems of specified category, strictly monolingual analysis and generation dictionaries, and transfer dictionaries based on language-pair-specific information.**
- **MT systems are linked to electronic dictionaries. Such electronic dictionaries can be of immense help even if they are supplied or used without automatic translation of text.**

Morphological Generation

- The aim in morphological generation is to produce the inflected form of a word according to the features and values in the Feature Structure.
- It is also necessary to reuse the linguistic resources created for analysis purpose.
- From a practical point of view, morphological generation is the inverted form of analysis, namely the process of converting the internal representation of a word to its surface form.
- A morphological generator designed for Tamil needs to tackle the different syntactic categories such as nouns, verbs, postpositions, adjectives, adverbs etc. separately, since the addition of morphological constituents to each of these syntactic categories depends on different types of information.

Morphological Generation

For example,

Root: MOUSE category (PartOfSpeech) : Noun Number:Plural

Stem: MOVE category (PartOfSpeech) : Verb Tense:Past

then morphological generation would convert these to the character strings

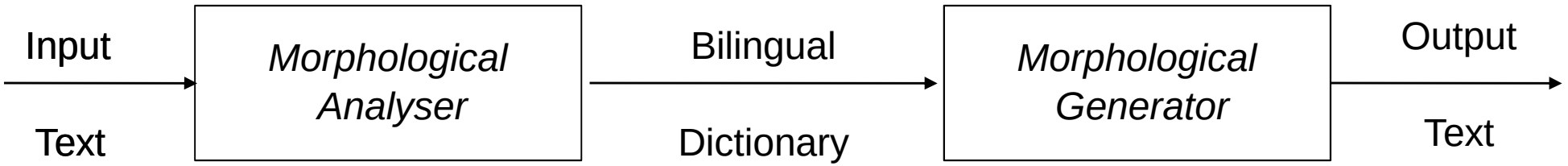
***mice* and *moved*.**

Suffix Joining Method

The identified suffixes are used along with the morphophonemic rules and morphotactic for developing the morphological generator.

While going from Malayalam to Tamil, there are about 11 different forms for a single stem in Tamil. But here only 6 forms are generated.

Implementation



Malayalam Input

Tamil Output

Algorithm

Step 1: Get the word to be analyzed.

Step 2: Check whether the entered word is found in the Root Dictionary.

Step 3: If the word is found in the dictionary, stop;

Else

Step 4: Separate any suffix from the right hand side

Step 5: If any suffix is present in the word, then check the availability of the suffix in the dictionary.

Then

Step 6: Remove the suffix present and then re-initialize the word without the identified suffix and go to Step 2.

Step 7: Repeat this process until the Dictionary finds the root/stem word.

Step 8: Store the Malayalam root/stem word in a variable and then get the corresponding Tamil word from the bilingual dictionary

Step 9: Check what all grammatical features does the Malayalam word have given and then generate the corresponding features for the Tamil word.

Step 10: Exit.

Result

Morphological Analyser Output:-

The entered word is kaaNippikkunnu

[Linkmorph] --------- kk , i

[Stem] --------------- kaaN

[Present Tense] ----- unnu

[Causitive] ---------- ppi

Morphological Generator Output:-

The generated forms of the Tamil word

kaaNukiRaan , kaaNukiRaar

kaaNukiRaaL , kaaNukiRaar

kaaNukinratu , kaaNukinRana

Conclusion

- **A significant part of the development of any machine translation (MT) system is the creation of lexical resources that the system will use.**

- **The proper functioning of a morphological generator necessitates efficiency in the generation of a word, once provided it roots or Stem and the corresponding feature values.**

Reference

- Ritchie, Graeme. 1985. *The Lexicon*. In Whitelock et al. (eds.), p. 225-256.

- D. Arnold, L. Balkan, S. Meijer, R. L. Humphreys, and L. Sadler. 1994. *Machine Translation: An Introductory Guide,* ch.5. UK: NCC Blackwell.

- Gülsen Eryioit and Esref Adalý. 2004.An Affix Stripping Morphological Analyser for Turkish. In *Proceedings of International Conference on AI and Applications*, Innsbruck 299-304. 16-18 February

- Rajeev R,R, Rajendran N and Elizabeth Sherly. 2008. A Suffix Stripping Based Morph Analyser for Malayalam Language, In *Proceedings of 20th Kerala Science Congress*, p 482-484, 28- 31 January.

- F.Och and H.Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1): 19-51.

- G.Grefenstette. 1998. The Problem of Cross-language Information Retrieval. In: G Grefenstette, ed. *Cross-language Information Retrieval.* Kluwer Academic Press, pp.1-9.

ANY QUERIES ??

THANK YOU