11-712: NLP Lab Report

Kartik Goyal

April 26, 2013

Abstract

This work aims to build a morphological analyzer for the Spanish language using Finite State Transducers(FST). The common rules for inflections of various base forms are inspected and are encoded in the FST. This method is quite effective as FSTs are able to handle a lot of regular rules and adding rules or edition of the analyzer becomes very comvinient by using operators like composition of FST, union of FSTs etc. The analyzer is built using two 1000-type corpora as the development set. The analyzer is also run on a 10000 word corpus.

1 Basic Information about Spanish

Spanish is a Romance(Ibero romance group) language that originated in Castile region of Spain. Apart from the other Romance languages, it enriched it's vocabulary from Basque and Arabic. Spanish is closely related to other Iberian Languages like Italian and Portuguese with which, it has 82% and 89% lexical similarity respectively. Spanish is written in Latin script. The interrogative and exclamatory clauses are introduces with inverted question and exclamation marks.

It is the second most spoken language by number of native speakers and is one of the 6 official languages of the united nations. It is spoken by around 329 million people all over the world. Spanish is the primary language of 20 countries worldwide. Apart from Spain, it is official language of many Latin American regions(Argentina, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Uruguay, Venezuela, Peru,Puerto Rico). It is also, the official language of Equatorial Guinea in Africa.

There are some grammatical and lexical differences between the Spanish spoken in regions of Spain and the Spanish spoken in Latin America. The main grammatical variations between dialects of Spanish involve differing uses of second person pronouns. For eg. In most of Spain, 'ustedes' and 'vosotros' are used according to the degree of formality, but only 'ustedes' is used in Latin America. There are some important vocabulary differences too between the dialects.

Spanish has an S-V-O grammar and is a heavily inflected language. It has a two gender noun system and the inflections of nouns, adjectives and determiners are caused by the number and gender. There are about 50 conjugated forms per verb which are caused by tense, number, person, T-V distinctions(formal) for second person, mood, aspect and voice. The modifying constituents tend to be placed after their head words. Also, usually the adjectives are placed after nouns.

Spanish is a morphologically rich language and this tool aims to carry out a resonably accurate morphological analysis of the language.

2 Past Work on the Morphology of Spanish

Tzoukerman and Liberman discus about a transducer which is essentially is a finite directed graph with edges labelled by relations between surface and lexical forms.

Santiago Rodriguez and Jesus Carretero describe the formal approach behind their spanish morphological analyzer, COES.

Atserias et al. discus about the morphosyntactic analyzer which can be executed in GATE environment.

Carmona et al. discus about MACO+, a tool for morphological analysis of spanish.

Sidorov et al. have developed a spanish morphological analyzer and have made their wordlist publicly available.

3 Available Resources

The reference grammar being used for this project has two sources:

- 'A Comprehensive Spanish Grammar'- Jacques de Bruyne
- 'A Reference Grammar of Spanish'- R.E. Batchelor, Miguel Angel San Jose

The lexicon I am using is derived from the words generated by the morphological analyzer page at `http://www.cic.ipn.mx/~sidorov/agme/`.

The corpora to test the system performance, have been built from the Judicial weekly and its Gazette provided on the official site of 'The Supreme court of Justice of the nation' from mexico `http://www.scjn.gob.mx/Paginas/Inicio.aspx`. The file used for building the corpus was the 'September 2011' issue of the gazette.

These judicial proceedings contain clean data and are dominated by the formal official language. But, this data also contains quoted statements by authorities and people involved in the proceedings. Hence, some personal words are also expected in the corpus. All the numbers and words containing single letters like 'a' were removed as they are not interesting from the morphological point of view.

A total of 23,000 types were found in the corpus. Corpus A and Corpus B have 1000 types each. The test corpus has 10,000 words.

This corpus is expected to be a fairly clean and diverse because apart from being an official record, it contains people's personal statements and opinions.

4 Survey of Phenomena in Spanish

Spanish is a morphologically rich language. The verbs, nouns, pronouns and adjectives demonstrate inflections.

- **VERBS:**

  Handling the verb inflection is extremely important because they demonstrate extremely rich inflectional phenomena. The verbs undergo inflection according to following categories:

  - Tense: Present, Imperfect, Preterite, Future, Perfect
  - Number: Singular, Plural
  - Person: First, Second, Third
  - Mood: Indicative, Subjunctive, Imperative, Conditional
  - Aspect: Perfective aspect, Imperfective aspect
  - Voice: Active, Passive

  The regular verbs can be classified into 3 categories based on their endings:

- -ar ended- Most frequent
- -er ended- Fairly frequent
- -ir ended- Least Frequent

All the regular verbs are conjugated in a standard manner. All the verbs also havew gerund and past-participle forms The perfective aspects involve inflection of the auxilliary verb 'haber', which is attached to the past participle form of the root. These phenomena can be accounted for by focusing on the inflections of 'haber' alone. Irregular verbs have slightly different inflection rules than those of the regular verbs. Some of the irregular verbs are extremely frequent and imortant in Spanish namely 'Haber', 'Tener', 'Ser' and 'Estar'. Other irregular verbs either change the stem's spelling or inflect in a different manner than the regular verbs. The radical changing verbs like 'pensar', 'holgar' change $e \rightarrow ie$ and $o \rightarrow ue$, when they inflect for present indicative and present subjunctive cases. Some verbs change $e \rightarrow i$, $o \rightarrow u$ etc. In some verbs, which end in -iar and -uar, the i or u is stressed in the present indicative and the present subjunctive forms. The diphthong in the stem of some infinitives becomes two syllables with the stress on the second syllable, when the stem of the verb is stressed. Many verbs ending in '-cer'(Nacer) and 'cir' inflect for prsent indicative singular form by ending in '-zco'.In a number of verbs(eg. Secar), spelling of stems ending in 'c','g','gu','qu' and 'z' has to change before ending beginning with 'e' in order for the pronunciation of the stem to be preserved.

- **NOUNS:**

  The nouns inflect according to 2 genders X 2 Number cases. Nouns ending in -o are masculine, with the only notable exception of the word mano ("hand"); -a is typically feminine, with notable exceptions. The nouns ending in 'cin' are generally feminine. A small set of words of Greek origin and ending in '-ma', '-pa', or '-ta' are masculine. Many nouns have same spellings for both female and male forms, thereby, making a deterministic conclusion about the gender of a noun difficult.

  Plurals of nouns in Spanish are formed in the following way:

  When the noun ends in an unstressed vowel, '-s' is added.

  When the noun ends in a consonant, a stressed vowel '-es' is added.

  Nouns ending in 'z', change the 'z' to 'c' before 'es' in plural.

- **ADJECTIVES:**

  Adjectives too inflect according to 2 genders X 2 Number. The adjectives ending in 'o' are male and their female forms end in 'a'.

  Adjectives ending in -ete,-ote are male and their female forms have the 'e' replaced by 'a'.

  Adjectives ending in '-an', '-in' and '-on' are male and their female forms have a appended at the end.

  The adjectives inflect with number in exactly the same way as nouns inflect with number.

  The adjectives also inflect according to the degree of comparison.'-simo' at the end and 're-','rete-', and 'requte-' at the beginning are used for expressing a superlative quality.

  Finally some adverbs are generated from adjectives by adding a suffix '-mente' to them. This is a very common phenomenon.

- **ARTICLES:**

  Both Definite and Indefinite articles inflect with 2 gender X 2 Number. They have a fixed form without any variations.

  However 'de' and 'el' are combined to 'del' and 'a' and 'el' are combined to 'al'.

- **PRONOUNS:** They are seldom used but they too have fixed inflections with 2 gender X 3 Person. Interestingly, the clitic pronouns, which can be direct or indirect, generally follow the infinitive form of the verb(eg. 'damelo').

5 Initial Design

'FOMA' and 'lexc' were the tools used to build the morphological analyzer. This round of development focuses primarily on the nouns , adjectives and verbs. The lexicon, containing base forms, used is extracted from the list of words used by Sidorov et al. in their morphological analyzer.

- **NOUNS:** The nouns are divided into these categories:
  - Nouns ending in -o: These are masculine and inflect according to number.
  - Nouns ending in -a: These are feminine and inflect according with number.
  - Nouns ending in -cin: These are feminine. Interestingly, their plural forms end in -en and the accent on is removed as it becomes the second last syllable which is emphasized by default in speech.
  - Nouns having other endings: They have -es plural endings and their gender cannot be determined unless we know of their origins.
  - A specialized list of nouns which have -ma, -pa, -ta endings was used as these nouns are masculine inspite of ending in a.

  The plural forms of the nouns ending in -z, end in -ces instead of -zes. Various rules were written to handle the irregular cases mentioned above.

- **ADJECTIVES:** The adjectives are divided into these categories:
  - Adjectives ending in -o: They were treated similarly to nouns, but an addtional case is handled. The adjectives ending -simo always denote the masculine superlative forms of the adjectives, which was handled by appropriate rules.
  - Adjectives ending in -a: Similar to previous case, adjectives ending in sima are feminine superlative forms.
  - Adjectives Having other endings: These adjectives also were handled in a manner similar to the nouns. But, care was taken to handle the adverbs which are derived by inflecting the adjectives. Generally, -mente suffix is added to the feminine form of an adjective to convert it to an adverb.

  The pluralization rules are exactly similar to the rules for nouns.

- **VERBS:** The verbs are divided into these categories:
  - Verbs ending in -ar
  - Verbs ending in -er

– Verbs ending in -ir

All the regular rules for the verb inflections described in the previous section were implemented for all the categories of verbs. Among the irregular inflections, following are handled:

– 'Haber', 'Tener', 'Estar' and 'Ser' forms were hard-wired.
– Verbs that change the stem from $e \rightarrow ie$, $o \rightarrow ue$, $i \rightarrow ie$, $u \rightarrow ue$, $e \rightarrow i$, $o \rightarrow u$ in the present indicative and subjunctive forms, were handled.
– The verbs ending in -cer and -cir, have a -zco ending for their present singular indicative form.

The articles and the pronouns were not hard-wired. Also, the clitics were not handled.

6 System Analysis on Corpus A

My informant, Jonathan Barker, manually evaluated the analyzer's performance on Corpus A. Since, guesses are also involved in the analyses for words not in the lexicon, the evaluation metric defined was simple and lenient. The recall of the system was calculated and guesses were evaluated to be correct if they managed to yield one or more correct analyses. The average number of guesses the system produces for a word is 5.

The recall on Corpus A with the above-mentioned metrics was **79.8 %**.

7 Lessons Learned and Revised Design

Based upon the feedback given by my informant, following shortcomings of the analyzer were listed:

- Clitics like 'damelo, which involve verbs following direct/indirect pronouns.
- Implementing basic infinitive forms of verbs.
- Adjectival use of Past-Participles.
- Verb inflections involving diphthongization and accent modification.
- Other irregular forms which rely on lexicons.
- Closed forms groups like prononuns, articles, prepositions have not been implemented yet.

8 System Analysis on Corpus B

Before analysis on corpus B, the following improvements were added:

- Implementing basic infinitive forms of verbs.
- Closed forms groups like prononuns, articles, prepositions have been implemented.

The number of types guessed reduced in this corpus.

9 Final Revisions

An analyzer has been improved to handle a class of enclitics. These clitics satisfy the rule of pronouns attaching as suffixes to the verb forms. Finally, to improve recall, the variants of words involving accented vowels are also checked and their analyses are returned accordingly. This assumption is a good assumption because as a rule, if the vocal emphasis is on the second last syllable, then the vowel is not accented. But, because of inflections and the constancy of the vocal emphasis, accents are introduced in the spelling. The final normalization rule would account for all such inflections. In case the analyzer is guessing, this rule would also result in an increase in the number of guesses as both the accented and the non accented forms will be present in the analysis.

The analysis of the analyzer on a 10000 word corpus can be found in the folder 'results', having name 'Corpus-C_analysis'.

10 Future Work

This morphological analyzer in its current state can be improved by incorporating larger lexicons so that it resorts to guessing less often. Inflections involving modifications in the surface forms due to diphthongization need to be handled in an elegant manner. Currently, the analyzer handles a subset of cases where such inflections are prevalent.