INTRODUCTION

• Automatic syllabification is the process of determining the proper placement of syllable boundaries in a given word. For example, input: syllables — output: syllabifies.

• Syllables have been used as key features in text-to-speech (TTS) systems for many languages. Rule-based approaches have traditionally been the preferred syllabification method.

• Data-driven methods have been shown to perform significantly better for syllabification of English words (Marchand, Adsett & Damper, to appear) — perhaps because English is a language with a complex and irregular syllable structure (Bruck, Treiman & Carovolas, 1995).

• In contrast, Italian is considered to exhibit simple syllable structure (Seymour, Ari & Erkens, 2003).

• Research Question: Which rule-based or data-driven methods for automatic syllabification perform best in Italian, a language with simple syllable structure than English?

ITALIAN LEXICON

The Italian lexicon used in this research is part of the Italian Festival TTS project (Cosi, Tesser, Guettler & Avesani, 2001). It was created by TTOI (Istituto Trentino di Cultura – Instituto per la Ricerca Scientifica e Tecnologia) ISTC-SPFD CNR (Istituto di Scienza e Tecnologia della Cognizione – Sezione di Padova “Fenitico e Didatticaletica” – Consiglio Nazionale della Ricerca) and is freely available at www.gd.istc.cnr.it/software/IT-Prima/tar/0/ITEx_Italian.tar.gz

After preprocessing, the resulting lexicon (referred to as the Full Italian lexicon) consisted of 44,720 entries.

RULE-BASED ALGORITHMS

• Cioni’s algorithm for the syllabification of written Italian

Cioni (1997) presents a method using what he claims to be a “minimal set of rules” developed with the assistance of Italian linguists. Source code (C) is available from: www.di.unipi.it/~lcioni/AltroSoftware/sillabatore.tar.gz

A subset of the rules are shown to the right.

• Hall’s ordered rules for Italian syllabification

Hall (1974) lists six ordered rules for breaking single Italian words into syllables (shown below).

• Using these rules, a rule-based automatic syllabification program was implemented for comparison against the other methods.

1. CVC -> CV C
2. VC CV -> VC C
3. CV C -> C V C
4. VCV -> VCV
5. CVVC -> CVVC
6. VCV -> VC V, VC V, VCV, VC V, VC V
7. VCVC -> VCVC
8. CVCV -> CV C V
9. CCVC -> C V C
10. VCVC -> VCVC

• Beramini’s SYL-LABE program

Bergamini’s rule-based syllabification algorithm is called SYL-LABE (MacKinney-Romero & Goddard, 2006) and was implemented in C.

• It is available for download from: www.placetofly.it/pages/express/c/Utility/

• One sample rule used in this system is: VCV -> VC V

RULE-BASED ALGORITHMS (cont’d)

Conclusions from the experiments include:

1. The data-driven methods (SbA, SYLL-LABE) have significantly better performance on the whole Italian lexicon than the rule-based methods (Syllabification, Cioni).

2. There is an overlap between the rule-based and data-driven methods.

3. The difference between the best rule-based method and SbA is still significant.

4. Data-driven methods have been shown to perform significantly better for syllabification of Italian lexicons.

5. Syllables have been used as key features in text-to-speech (TTS) systems for many languages.

6. Although Italian results are not as strong, the difference between the best rule-based method (SYL-LABE) and SbA is still significant.

7. Overall performance on Italian is higher than on English.

REFERENCES


ACKNOWLEDGEMENTS

The authors wish to thank Piero Cosi for providing help with the Italian lexicon. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the National Research Council (NRC) Graduate Student Scholarship Supplement Program (GSSSP), and an Izaak Walton Killam Predoctoral Scholarship.

RESULTS

Three rule-based Italian syllabification methods were compared to two data-driven ones, revealing that data-driven methods perform best.

From comparison with English results, it is evident that, regardless of the method used, performance on the Italian lexicon is significantly better. This confirms beliefs that Italian is more straightforward to automatically syllabify than English.

The best data-driven method (SbA) reaches a word accuracy of 97.70%, whereas the best rule-based method (SYLL-LABE) achieves 89.77% on the Full Italian Lexicon.

These results suggest that, when a syllabification procedure is included as a component of a TTS system, a data-driven method is a more appropriate choice than a rule-based approach, even for languages with low syllabic complexity.