Determining the gender of the unseen name through hyphenation.

Robert H. Warren[1] and Christopher Leurer[2]

[1] University of Waterloo
Waterloo, On, Canada
rhwarren@uwaterloo.ca
[2] McGill University
Montreal, PQ, Canada
leurer@math.mcgill.ca

Abstract. The accepted method of determining name gender is to use a probabilistic model based on observations, which fails to classify unseen names. We attempt to solve this by utilising a hyphenation-driven method which is also more space efficient.

The ability to cross-check several fields within a record is of value as it permits us to validate the information provided. We concentrate here on determining the probable gender of a name so that it can be compared with other gender-related fields. Thus, a record of a person with the salutation "Mr.", given name "John", and whose gender is coded as a woman may be of questionable value and require additional inspection.

A common method involves the use of a probabilistic model built from name observations [1, 2]. This approach suffers from the inability to provide information on names which have not been previously observed. Alternatives which have been explored are the use of edit distance methods [3] or soundex matching to identify similar names. In our method we use the hyphenated form of the name to infer its gender. [3]

We used a trivial method to generate rules from hyphenated words by extracting the last token and using this as the basis for a probabilistic model (e.g.: "elizabeth" would hyphenate to "eliz-a-beth" from which we would extract the "beth" suffix.). The generated rules can be looked up without hyphenating the names themselves by matching the right-hand-side of the word to the rules. For this experiment, we used the readily available LaTeX hyphenation files for the English and German languages along with an open-source hyphenator [4, 5]. The hyphenation method is not critical as it provides a segmentation method; syllables could also be used, but with a high complexity cost.

A dataset of name-gender pairs generated from GEDCOM [6] genealogy files with over 60,000 individuals and more than 5,000 unique names was used to validate the method. To avoid cumbersome data cleaning and character set issues, we only processed names which contained the basic US-ASCII character set.

Table 1 contains a breakdown of the precision and recall figures for each of the classification models. We found that the hyphenation driven classification was correcly

[3] The authors wish to thank Dr. Brett Kessler of Washington University in St. Louis for help with this approach.

assigning gender in 80% of applicable cases. Interestingly, only about 20,000 names were required for all models to return a consistent performance.[4]

Method	Precision	Recall	Decision table size in rows
Name lookup	85%	87%	5742
Hyphenation lookup	87%	96%	1560
Name + Hyph. fallback	93%	96%	7302
Hyphenation (unseen only)	80%	10%	1560

Table 1. Precision / Recall measures.

The hyphenation model is very efficient as it requires 66% less rules than a name lookup model with a comparable performance on observed names. Hyphenation was able to classify an additional 10% of the names with a high precision. About 3% of names remained unclassifiable.

A valuable aspect of using a hyphenated model to identify gender is that the recall histogram of its rules is narrower than a basic name model. In situations where space is very limited, such as in field data-entry applications, a hyphenated model delivers a higher value than a standard name lookup model. This novel heuristic to classify unseen names is computationally inexpensive and allows us to cross check database records for proper gender identification.

References

1. F. Patman and P. Thompson, "Names: A new frontier in text mining," in *Proceedings of the First NSF/NIJ Symposium on intelligence and security informatics* (H. C. et al., ed.), (Tucson, AZ), pp. 27–38, Springer-Verlag, June 2003.
2. E. Guy, *Parserat*. Internet, 2004.
3. M. Bilenko and R. J. Mooney, "Learning to combine trained distance metrics for duplicate detection in databases," Tech. Rep. Technical Report AI 02-296, Artificial Intelligence Laboratory, University of Texas at Austin, Austin, TX, Feb. 2002.
4. F. M. Liang, *Word Hy-phen-a-tion by Com-put-er*. PhD thesis, Department of Computer Science, Stanford University, Stanford, CA 94305, August 1983.
5. D. Tolpin, *TeX Hyphenator in Java*. 2003. `http://www.davidashen.net/`.
6. F. H. Department, *The GEDCOM Standard Release 5.5*. The Church of Jesus Christ of Latter-day Saints, January 1996.

[4] The complete probabilistic name and hyphenated models can be found on the web at `http://jill.math.uwaterloo.ca/~warren/name-gender.xml` and `http://jill.math.uwaterloo.ca/~warren/h-name-gender.xml` respectively.