Converting Unicode Lexicon and Lexical Tools for ASCII NLP Applications

Chris J. Lu, Ph.D.[1,2] and Allen C. Browne [1]
[1]National Library of Medicine, Bethesda, MD; [2]Lockheed Martin/MSD, Bethesda, MD

Abstract
The NLP SPECIALIST Lexicon and Lexical Tools, distributed by National Library of Medicine (NLM), have been released in Unicode (UTF-8) format since 2006. Lexicon is used as corpus while Lexical Tools are used as software packages in NLP (Natural Language Processing) projects. Some NLP projects still only deal with ASCII (7-bit) characters. This paper describes how to convert UTF-8 Lexicon and integrate Lexical Tools to a pure ASCII NLP project, MetaMap.

1. Introduction

Converting Unicode text, corpora, NLP software packages to ASCII are common processes in ASCII NLP projects. This task is challenging because: 1) ASCII conversions are not one-to-one mapping. 2) Some Unicode characters are misused before the conversion. For example, μ (mu, U+03BC) and µ (micro sign, U+00B5) are commonly misplaced because of their similar graphical representations. 3) Wrong conversions occur when the converted ASCII word does not exist or has a different meaning. For example, the French borrowing "divorcé" means a man who is divorced. This word has no pure ASCII spelling variant in Webster's Dictionary, while the converted ASCII word, "divorce", is another closely related word. To convert corpora and NLP packages to ASCII is even harder. It usually requires knowledge from domain experts and linguists for accurate conversions. The following sections describe our approaches and results of converting Lexicon and integrate Lexical Tools to an ASCII NLP project.

2. Converting the Corpus, Lexicon

The NLM Lexical Systems Group generates ASCII Lexicon from Lexicon to support NLP projects only dealing with ASCII. The algorithms are: 1) Convert lines containing non-ASCII characters in Lexicon to ASCII by using Lexical Tools - ToAscii [1]. 2) Delete lexical records if the converted citation is not known by Lexicon or has a different meaning. Theoretically, the ASCII Lexicon is a subset of Unicode Lexicon since ASCII is a subset of Unicode. ASCII converted records are deleted if the converted citations are not known to (contained inside) Lexicon. For example, the record of "Müthing" [E0573093] is deleted because its ASCII conversion, "Muthing", is not in Lexicon. 3) Remove the ASCII conversions of spelling variants if they are duplicated. For example, "résumé" (a spelling variant of "resume") is removed because its ASCII conversion, "resume", is a duplication of the citation. 4) Delete conversions if the meaning changes. For example, "µm" (a spelling variant of "mu" [E0041164]), is deleted because its ASCII conversion, "mum" [E0041369], has a different meaning. Table-1 shows numbers of all four cases in the ASCII conversion for Lexicon, 2010-11.

Release	Case 1	Case 2	Case 3	Case 4
2010	4,345	29	3,685	268
2011	5,801	42	4,906	377

Table-1 ASCII conversion details for Lexicon

3. Converting the NLP Package, Lexical Tools

Lexical Tools uses Lexicon to generate nine relational database tables for various lexical variants permutations. The ASCII version of Lexical Tools can be derived by generating these database tables in ASCII from ASCII Lexicon and then reloading them to the database of Lexical Tools. However, this traditional approach is tedious and not practical for end users. A much easier approach is to implement an interface to the outputs from Lexical Tools for pure ASCII applications. The algorithm of the interface includes: 1) Convert the results to ASCII by using Lexical Tools – ToAscii. 2) Remove the results if not known to Lexicon. 3) Remove the results if the ASCII conversion is duplicated.

4. Application and Conclusion

MetaMap only deals with ASCII and is used to map biomedical text to concepts in the UMLS Metathesaurus. The ASCII Lexicon and ASCII conversion interface of Lexical Tools are used in MetaMap to retrieve citation forms, spelling variants, inflectional variants, derivational variants, and normalizations. A test suite is developed to compare the results from both approaches in the previous section. The interface approach is easy and generic and provides identical results of the traditional approach over 0.5M test cases for 2010 release.

References

1. Lu, Chris J.; Browne, Allen C.; Divita, Guy, "Using Lexical Tools to Convert Unicode Characters to ASCII", Proceeding of AMIA 2008 Annual Symposium, Nov. 8-12, 2008, Washington DC, p. 1031