Using Lexical Tools to Convert Unicode Characters to ASCII

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

Abstract

Unicode is an industry standard allowing computers to consistently represent and manipulate text expressed in most of the world's writing systems. It is widely used in multilingual NLP (natural language processing) projects. On the other hand, there are some NLP projects still only dealing with ASCII characters. This paper describes methods of utilizing lexical tools to convert Unicode characters (UTF-8) to ASCII (7-bit) characters.

1. Introduction

The SPECIALIST Lexical tools, 2008, distributed by National Library of Medicine (NLM) provide several functions, called LVG (Lexical Variant Generation) flow components, to convert Unicode characters to ASCII. In general, ASCII conversion either preserves semantic and/or graphic representation or facilitates NLP. Different NLP applications might apply different methods for the ASCII conversion due to different requirements and objectives. There is no single standard method for ASCII conversion. For example, character, ™, can be converted in the following ways:

- Graphic: TM
- Semantic: ![TRADE MARK SIGN]!
- Graphic and Semantic: (TM), or (tm)
- NLP: empty string, consider ™ as a stopword

2. Methods

The Lexical tools provide five types of methods for ASCII conversion. They are detailed below:

2-1. Unicode normalization

The Unicode standard allows some characters to be described as a combination of an ASCII character and a diacritic mark. Non-ASCII diacritic and ligature characters are common used in Spanish, French, and English documents. An LVG flow, -f:q, strips the diacritic from such characters. Unicode also allows ligature characters to be described as a combination of two ASCII characters. Another LVG flow, -f:q2, splits these ligature characters into their respective ASCII parts.

2-2. Table lookup mapping

In general, table lookup mapping method is applied to all Unicode characters are not handled by Unicode normalization algorithm (described in Sec. 2.1). Unicode symbols and punctuation are very confusing not only because they look alike, multiple defined (in different Unicode blocks), but also because text editor software automatically change them during the editing and transaction. LVG flows, -f:q0 and –f:q1, are used to preserve the semantic and/or graphic representations in the ASCII conversion for Unicode Symbols and characters respectively by using this table lookup mapping method.

2-3. Recursive algorithm

Some Unicode characters require multiple steps, combining the methods described above, in ASCII conversion. A recursive algorithm is implemented in LVG flow, -f:q7, for this purpose.

2-4. Table lookup mapping and strip

Some Unicode characters do not belong to categories mentioned above. A local table of conversion values is used to convert these to an ASCII representation. For example, Greek letters are converted into fully spelled out forms. 'α' is converted to "alpha". Non-defined Unicode characters, (such as ™, ©, ®, etc.), are treated as stopwords and stripped out completely to ensure pure ASCII conversion. This table lookup and strip function is represented by LVG flow, -f:q8.

2-5. Pure ASCII conversions

In addition to above fundamental LVG flows, Lexical Tools provide more sophisticated flows to convert Unicode to pure ASCII, such as –f:q5, -f:q6, -f:N, -f:N3, etc.. The combined serial LVG flow, **-f:q7:q8** is the most powerful method and commonly used for pure ASCII conversion. All these flows can be configured according to user's specifications. Table 1 shows examples of ASCII conversion for LVG flows in Lexical tools.

LVG Flow	Input (UTF-8)	Output (ASCII)
-f:q	Déjà Vu	Deja Vu
-f:q0	"Quote"	"Quote"
-f:q1	⅔	2/3
-f:q2	spælsau	spaelsau
-f:q5	UMLS®	UMLS![REGISTERED SIGN]!
-f:q7	Ǣ	AE
-f:q8	α	alpha
-f:q8	Zadaxin™	Zadaxin
-f:q7:q8	UMLS®	UMLS

Table 1. Examples for ASCII conversion of LVG flows in Lexical Tools

The detail documents and examples on Unicode handling of Lexical Tools are available at following URL: http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/designDoc/UDF/unicode/index.html.

3. Conclusion

There are many different ways to convert Unicode characters to ASCII. The SPECIALIST Lexical Tools, (2008), provide various powerful methods for ASCII conversion and allow users to configure the tools to their specifications.