Text Normalization

Notebook:	Stanford NLP Book		
Created:	22-07-2020 23:29	**Updated:**	22-07-2020 23:54
Author:	kopalsharma2000@gmail.com		

At least three tasks are commonly applied as part of any normalization process:
1. Tokenizing (segmenting) words
2. Normalizing word formats
3. Segmenting sentences

Tokenizing words
UNIX COMMANDS
tr - tokenize the words every sequence of non alphabetic character to new line
-c ---> complements to non-alphabet
-s ---> squeezes all sequences into a single character
-n ---> to sort numerically than alphabetically
-r ---> reverse order that is highest to lowest

A tokenizer can also be used to expand clitic contractions that are marked by apostrophes, for example, converting what're to the two tokens what are

Tokenization is thus intimately tied up with named entity detection, the task of detecting names, dates, and organizations

In practical life, since tokenization needs to run before any other language processing it needs to be very fast.

Byte-pair Encoding
The intuition of the algorithm is to iteratively merge frequent pairs of characters
First Iteration

dictionary	vocabulary
5 l o w ,	d, e, i, l, n, o, r, s, t, w
2 l o w e s t	
6 n e w e r	
3 w i d e r	
2 n e w	

Second Iteration

dictionary	vocabulary
5 l o w ,	d, e, i, l, n, o, r, s, t, w, r
2 l o w e s t	
6 n e w e r	
3 w i d e r	
2 n e w	

Wordpiece
The wordpiece algorithm starts with some simple tokenization (such as by whitespace) into rough words, and then breaks those rough word tokens into subword tokens. The wordpiece model differs from BPE only in that the special word-boundary token appears at the beginning of words rather than at the end, and in the way it merges pairs. Rather than merging the pairs that are most frequent,

wordpiece instead merges the pairs that minimizes the language model likelihood of the training data. Basically the token that will give the corpus the highest probability.

Maximum Matching
The maximum matching algorithm (Fig. 2.13) is given a vocabulary (a learned list of wordpiece tokens) and a string and starts by pointing at the beginning of a string. It chooses the longest token in the wordpiece vocabulary that matches the input at the current position, and moves the pointer past that word in the string. The algorithm is then applied again starting from the new pointer position.

Sentencepiece
The SentencePiece model works from raw text; even whitespace is handled as a normal symbol.
Thus it doesn't need an initial tokenization or word-list, and can be used in languages like Chinese
or Japanese that don't have spaces.

Normalizing Word Formats
Word normalization is the task of putting words/tokens in a standard format, choosing a single normal form for words with multiple forms like USA and US or uh-huh and uhhuh.

Case Folding
Case folding is another kind of normalization. Mapping everything to lower case means that Woodchuck and woodchuck are represented identically, which is very helpful for generalization in many tasks, such as information retrieval or speech recognition. However this is not helpful in machine translation, sentiment analysis, text classification etc. as US and us have different meaning.

Lemmatization
Lemmatization is the task of determining that two words have the same root, despite their surface differences. The words am, are, and is have the shared lemma be; the words dinner and dinners both have the lemma dinner.

> Morphological Parsing
> A morphological parser takes a word like cats and parses it into the two morphemes cat and s, or a Spanish word like amaren ('if in the future they would love') into the morphemes amar 'to love', 3PL, and future subjunctive
>
> Stemming (Porter Stemmer)
> Naive version of morphological analysis is called stemming.

Segmenting Sentences
In general, sentence tokenization methods work by first deciding (based on rules or machine learning) whether a period is part of the word or is a sentence-boundary marker.