IE Case Study

Information extraction (IE) requires accurate pre-processing to mark up your text input stream. One such pre-processing area is Name Entity Recognition (NER) for person names, with the appropriate markup such as title, surname, given name(s), gender, etc. This helps immensely in further text processing such as co-reference resolution, subject/object identification, event sequencing and so on.

Let us consider gender identification. Within most cultures, intuitively, human beings are able to predict accurately the gender of a name (i.e., classify into “male”, “female” or “neutral”) – so we know that it’s possible to build a model for name gender. This has been illustrated many times for “Western” (English, French, Spanish and other romance languages) first-names, and we can assume we have accurate models of name gender for Western first names. We can also assume that Western surnames are gender neutral.

The typical Chinese names (in English) is more complicated than the Western case. For example, President Tony Tan Keng Yam has the title “President”, the Western first-name “Tony”, the surname “Tan”, and Chinese-given-name “Keng Yam”. While we know that the name as a whole is male, it would be useful to know whether “Keng Yam” by itself is male, female or neutral, in the same way that we are able to predict gender for Western first-names. We also note that, like Western names, Chinese surnames are gender neutral. Also, Chinese given-names need not be two words long; many are only one word. For example, the famous Chinese poet, Li Bai, has surname “Li” and given-name “Bai” while President Xi Jinping of China has surname “Xi” and given-name “Jinping” (corresponding to 2 Chinese characters). For simplicity, we will ignore the cases where Chinese names are longer than 3 words (in English).

We can assume that we have accurate tokenizers to do NER on a name, which is to extract the string from a text that is a name and to mark up the name components. Your tokenizer is able to distinguish Title, Western-first-names, Western-first-name-gender, surname (both Chinese and Western), and given-names (basically, the remainder after first-name and surname are identified).

So from the following sentences:

> President Tony Tan Keng Yam was the 7th president of Singapore. The first lady was Mary Chee.

the tokenized name markup will be the following:

> ((President)$_{\text{+Title}}$ (Tony)$_{\text{+first\_name,+male}}$ (Tan)$_{\text{+surname}}$ (Keng Yam)$_{\text{+given\_name}}$)$_{\text{+name}}$ was the 7th president of Singapore. The first lady was ((Mary)$_{\text{+first\_name,+female}}$ (Chee)$_{\text{+surname}}$)$_{\text{+name}}$.

You are a graduate student given the task to build a name gender model for Chinese-given-names in English, excluding titles, surnames (family names) and any Western first-names they may be using. You have very little funding for this so you know you have to be smart. You decide to build a supervised model to predict the gender of the given-names. You decide to crawl the internet to extract Chinese names from English news sources, obituaries, authored documents (journals, patents, etc.). In addition, you obtain the list of the top 1000 most common surnames in Chinese. From your 1-week crawl, you successfully obtain 10,000,000 names.

Answer the following questions:

a. Given that you need to come up with a list of Chinese given-names with their gender, propose how you would distinguish Chinese names from non-Chinese names in your list of 10,000,000 names?

b. Assuming that accuracy and quantity are trade-offs, is it more important to have more data, or more accurate data? Justify your answer.

c. You will notice that people do not have unique names, so your data is likely to contain multiple instances of the same given-names. Should you keep all instances of the given-names, or reduce your dataset to only unique given-names? Justify your answer.

d. Propose how you would establish an accurate ground truth for given-names gender in your data? Remember that you have three possible values for gender: male, female and neutral.

e. Do you need to validate your final data set? What problems might you expect to have, and how would you solve them?