Speekulate

Speech Kurdish Speculate Translation

Version 0.1

S. Ghafour

15 December 2011

Introduction

The Scope of this document is initially not strictly limited to a time-frame or deadline. It describes, in particular, a theoretical scenario and tests its potentials of implementation in a real world example.
As the outcome and the results depend on many factors, the scope of the delivered product can be modified during the process of development in an agile fashion.

Prelude

Speech to text, Kurdish and Speculate a translation will make Speekulate. The main idea is to use the cutting-edge technologies that are available at the moment to produce a challenging-creative product; it is machine challenging human. The applications of the end-product are unlimited with a dozen of positive side-effects.

Speekulate can be considered as, however it may remain as a theoretical research, multi-use translator; “Upload an English movie, play back with Kurdish subtitles.” This description sounds too simple and fragile for such a huge system. We should be aware of the fact that there is no such a product out-of-box and there is a demand on a an extesive research to create a creative mixture of software.

So why is it so important to have a system like Speekulate? Let us forget about Speekulate in terms of an end-product. While we will describe the steps of the process, it will become more obvious.

Intermezzo

The first important step to know is, the current status of the Kurdish language. In the following sections we will try to anatomize the situation from the very beginning.

The Kurdish Alphabet

In the late 90s of the previous century first Kurdish texts start to appear on Internet. Due to the lack of standards, imperfection of Unicode and lack of knowledge, Kurds were obliged to create their own fonts/letter-types. The Kurdish fonts were mainly based on Arabic or Persian structures. The glyphs in the font were simply edited and replace by Kurdish ones. The result was dramatic and disastrous, especially in terms of data-storage.

In the beginning of 2000 Kurditgroup created a standard of a digital Kurdish alphabet based on a correct Unicode and supported by a font that is not based on any other language. This standard by Kurditgroup is chosen unilaterally and can be formalized in a way to be adapted by the Kurdish government.

Sorani, is the *de facto* official dialect of the Kurdish Government. However, there are couple of issues that should be addressed by Kurdish linguists. That will create a wider support for this standard. Take a look at ک and ك. In terms of data-storage this is a very important issue to tackle.

First Step: Official normalization of the Kurdish alphabet and their corresponding codes as described in Unicode Tables. Without taking care of this step, we will create a monster that will be hard to kill in the future.

Word list of the Kurdish language

This Word list is crucial for Speekulate as well as to create a Spell check system. To produce such a list there is a big need to cooperate with Kurdish linguistic institutes and make it digitally available. The institutes can continuously update and improve the list. In the mean time it can provide documents about Kurdish morphology that are valuable to Statistical Machine Translation Systems.

Second Step: The official Kurdish Word list.

Creating an acoustic model

A phonetic transcription (needed in case of transforming Kurdish speech to text). This step can be skipped as the purpose of Speekulate is to create English to Kurdish translation system. However, having an acoustic Kurdish model can be useful in many ways and can be used in a wide range of applications.

Third Step (Optional): Kurdish Speech to Text.

Statistical Machine Translation

Creating such a system has multiple benefits. At this moment the most of existing systems do not support Kurdish due to the lack of sufficient Corpora, ex. Google Translator. This can be described as the most complicated and time consuming part. In this case a Shallow Translation System can be more effective.

Fourth Four: Machine Translation.

Choosing the systems

Most of the above described processes can be supported by one or more softwares that are freely available. In order to choose the best systems and train its behaviors there is a need of extensive research and high technical skills.

Glue the systems

In the final step the chosen systems will be brought together in an easy-to-use platform supported by a user interface/front-end.

Last Step: User interface.

The team

As mentioned before to build such a huge platform, it is required to have highly skilled and motivated people. The team can exist of Kurdish and non-Kurdish developers in and outside the country.

Time-frame

The lead-time of this project depends mainly on the speed that the different Kurdish organizations can operate. To have a better view on this, more research is needed.

Conclusion

The final result, Speekulate, depends on the quality of each step. I am confident that this project will address many problems of the Kurdish language. It creates an opportunity to bring different institutes together that are at the moment not participating in the new technologies era.