Error Types in Natural Language Processing in Inflectional Languages

1

Gregor Donaj
University of Maribor, Slovenia

Mirjam Sepesy Maučec
University of Maribor, Slovenia

INTRODUCTION

Natural Language Processing is the area of research that explores how humans can interact with computers in natural language. It is a subfield of Artificial Intelligence. The development of NLP applications is a challenging task, as natural languages are ambiguous, and not structured as strictly as programming languages, which were developed for the purpose of human-computer interaction. Today, we are faced with large amounts of data, available on the internet. The development of Natural Language Processing applications provides invaluable insight into very diverse data sources, which are not manageable by humans, but by the use of technology.

There are many different tasks in Natural Language Processing. To cover all of them within this article would be impossible. This article focuses only on two prominent Natural Language Processing applications: Automatic speech recognition and machine translation. The task of automatic speech recognition is for a machine to recognise the spoken words from an audio signal. This can be done on pre-recorded speech or live from a microphone. Essentially, it is the conversion from speech to text. The application area of automatic speech recognition is very broad. For example, it can be used for subtitling TV programmes for deaf people. Machine translation is the task of translating some text from one natural language to another. It can be applied to catch the meaning of a text written in a language that the reader does not understand.

The effectiveness of Natural Language Processing methods varies greatly, depending on the language under consideration. Highly inflected languages belong to a group of the most difficult languages to process. This article deals with one representative of this group, Slovene. It is a morphologically rich language with complex grammar, which harms the performance of Natural Language Processing tasks. The motivation of this article is to analyse errors that result from the characteristics of Slovene, and are not so frequent in Natural Language Processing of other languages.

BACKGROUND

Natural Language Processing

Large amounts of data are being created online every day, and many data are in the form of text in a natural language. The idea is to process and examine the data and to uncover new knowledge. Computers know how to process structured data, but data in natural language is unstructured. It requires specialised

DOI: 10.4018/978-1-7998-3479-3.ch006

approaches to process it. The research field that copes with this phenomenon is called Natural Language Processing.

Natural Language Processing (NLP) is a sub-field of Artificial Intelligence. It focuses on enabling computers to understand and process natural languages as humans do. Although NLP research has a long history, many problems are still unsolved. Computers are far behind human abilities. We still do not know precisely how humans process language. Despite that, today, computers offer many, quite useful applications that are based on natural language. The machine learning evolution spurred remarkable technology breakthroughs. NLP evolved from a time-consuming process where rules were handwritten by humans, to unsupervised learning, where computers learn from data by themselves.

There are many interesting tasks which are based on NLP:

- Document summarization: Automatically generating synopses of large bodies of text.
- Automatic speech recognition: Transforming voice into written text.
- Speech synthesis: Transforming text into voice.
- Machine translation: Automatic translation of a text (or speech) from one language to another.
- Sentiment analysis: Identifying the emotions and subjective opinions within large amounts of text.
- Semantic analysis: Interpreting human sentences logically.
- Natural language understanding: Transforming the meaning of a text into a structured semantic form.
- Natural language generation: Generating text from structured data in a readable format with meaningful phrases and sentences.
- Question Answering: Generating answers to questions in the form of a sentence. It is based on natural language understanding.
- Dialogue systems with virtual assistants: Using natural dialogue that mimics a live agent interaction.

Every day, new ideas for applying NLP arise. The goal of almost all NLP tasks is to take raw language input (in written or spoken form) and use linguistic knowledge and algorithms to deliver higher value to the user.

In the continuation of this article we will focus on two NLP tasks: Automatic Speech Recognition and Machine Translation.

Automatic Speech Recognition

The most natural, and also essential, form of human interaction is by speech. In the modern world, however, more and more of our interaction is not only with other people, but with information systems. In the pursuit of a natural interaction between humans and machines, Automatic Speech Recognition (ASR) systems have been developed for the better part of the 20th century in order to be complementary with traditional methods of human-computer interaction, e.g., keyboards and graphical user interfaces. Today, the advancement of speech recognition can already be observed in commercial applications, such as home assistants.

Other applications of ASR may be the transcription of spoken dialogue into text form for the hearing impaired and speech to speech translation, where speech recognition is performed first, then machine translation, and then speech synthesis, in order to enable spoken interaction between people who do not speak the same language.