# Tittle Page

一种文本相似度方法解决AIML聊天机器人中中英文匹配问题

A method based on text similarity to solve the Chinese matching problem of AIML chatbot

# Abstract

AIML是一种基于模板的聊天机器人的代表。但是它的匹配方式是基于树的结构来进行逐个词的匹配，这适应了英文的语法结构。英文句子自带分词。但是这种匹配方式不适应中文句子。这篇论文使用编辑距离这种文本相似度方法代替AIML聊天机器人的匹配结构，可以进行中英文的语料文本匹配。并提供了新的聊天机器人AIML变种框架，经过测试取得了好的效果。

AIML is an important kind of chatbot. But its matching mothed based on matching tree is design for English language. When it is applied to Chinese language. It causes many problems. The paper provides a simple method to solve the matching problem and keep the xml structure in the answer template. It performs well in our robot platform.

# Introduce

在智能机器人RIMA项目中，我们在实验室制造了一个可以与人交互和地图导航的机器人。我们需要一个可靠的聊天机器人来进行传达命令给机器人完成人机交互。由于实验室位于中国，同时拥有中英文语言使用者，所以我们需要一个聊天机器人可以同时满足中英文命令交互和简单聊天。

In RIMA intelligent robot project, it is a robot that can communicate with human and slam in the laboratory.

【AIML】

AIML是1991年xx教授发起的一个著名的聊天机器人开源项目，其中最著名的要数ALICE聊天机器人。并获得了多届聊天机器人最高奖项勒布那奖。2012年推出aiml2.0版本。至今AIML聊天机器人依然活跃在网络上，2019年的aiml组织推出xx的聊天机器人依然是aiml结构。xx聊天机器人现在是一名twitch主播。

AIML使用模板进行存储问题和答案，使用aiml标记语言来进行规则推理。在聊天机器人初始化后，所有问题会在内存中形成匹配树，逐个匹配每个输入的词与树中的节点，从而找到匹配的树，从而返回答案模板给解析器并进行推理和返回答案。匹配树可以加快搜索的速度而且可以很好的进行模板模糊匹配。但是这都依赖于英文句子自带分词。AIML可以通过通配符提取问题模板中的内容和匹配更多问题。

【Chinese problem】

但是aiml聊天机器人是根据英语语法特点设计的聊天机器人。因为英文语法自带分词，所以匹配树可以直接根据词生成节点。然而中文句子不自带分词，中文句子分词依然是个困难的问题。

其次相比英文语法，中文语法短句倒装句居多。比如英文句子“who get the apple”，中文表达方式有很多，“谁拿了苹果”，“拿了苹果的是谁”，“是谁拿了苹果”。英文中的一种表达方式，然而中文这三种表达都可以。

如果考虑到实验室实际需求中面临，中英文语言使用者的命令和简单聊天需要同时满足，所以这也带来了新的问题。

【相关研究】

[中国互联网的通用方法]

在中国工业界使用aiml构建简单的聊天机器人的时候，他们一般使用分词工具，对输入问题和候选答案进行分词。

一般有两种方式，对输入问题进行分词，语料库不分词，当语料库生成匹配树的时候进行分词。但这样显然每次生成聊天机器人将大量消耗内存和网络交互进行API交换。第二种方式，对输入问题进行分词，语料库分词并存储。但是这样语料库的处理更新和管理带来了大量问题。另外一方面是回答模板中并不仅仅包含返回的回答，可能包含一些歌曲名等信息，这些被回答模板解析器解析后发送给相应的程序处理，他们不能被分词。这样造成了回答中一部分分词一部分不分词，这会造成日后语料库管理的混乱和困难。

这种处理方式有两个问题，一方面中文分词的准确性一直是个困难问题，即使使用同样的分词工具，在几个月后同样的内容可能分词结果并不相同。另一方面，中文语法的多变性无法处理，及其相似的短句，AIML的书匹配无法判断他们是同样的问题。这是中文语法特点的所决定的。

[北京师范大学方法]

北京师范大学在开发聊天机器人的时候，在参考了中国工业界一般的AIML处理方法的基础上，进行分词存储语料库。并提出了基于词相似度的同义词处理来处理同义词问题。

使用主谓宾识别的方法解决了中文短语词序问题，这是一个好的方法。

但是这个方法在面对复杂性短语组成的语句中依然面临问题。



两者同样的意思然而难以通过主谓宾识别来进行调整成为同样一句话。



如图所示，句子成分识别依然存在问题。http://ltp.ai/demo.html

另外这样也丧失了AIML自带的通配符模板匹配的灵活性。

【chatterbot聊天机器人灵感】

chatterbot是一个机器学习的检索型聊天机器人框架，它使用机器学习方法来进行提取候选答案，经过贝叶斯过滤来解决time歧义和数学问题chatterbot使用编辑距离来处理问题和语料匹配问题。编辑距离可以处理多语言问题，也可以很好的处理多种语言的匹配问题。

【假设】

所以一个主意就是结合编辑距离这种机器学习匹配方法来进行匹配。上文说过匹配树可以大大加速匹配速度，这是当时受限于当年的电脑硬件性能。通过编辑距离可以很好的匹配相近的语料库句子，无需对中文进行分词。并且适合多语言。对于中文的短句倒装和多变的说法，只要主要的短语不变，编辑距离能够相对3较好的识别出来。

# Methods

A文本相似度

编辑距离文本相似度

B编辑距离文本相似度匹配

B新的聊天机器人结构

# Results

经过在网页和智能机器人测试。可以很好的支持中英文聊天。

可以看出聊天机器人可以很好的支持中文和英文的匹配，并支持相似度匹配，而且对于部分复合短语的倒装问题也有比较好的匹配率。

# Discussion

虽然编辑距离可以很好的解决中文聊天机器人匹配问题，并且适合多语言处理。但是匹配方式的改变也同样使得通配符匹配的功能丧失。这意味着聊天机器人将无法提取通配符所代表的实体。以及进行统配模板匹配。这个问题也同样出现在北京师范大学的聊天机器人中。

如果使用KMP不断尝试进行通配而且进行文本相似度的匹配的花，将极大的消耗匹配速度，它相当于把句子的每种顺序组合都尝试过一遍。这个代价太大了。

但是幸运的是实体抽取部分可以有语义理解模块来进行处理。可以使用语义理解聊天机器人和aiml聊天机器人组成多层次的聊天机器人来组成双层结构的聊天机器人弥补匹配方式的缺失。

# Reference

# Acknowledgement