Collaborative Arabic Morphological Service

Natheer Khasawneh Eyad Taqieddin Ahmad Al-Hammouri

Jordan University of Science and Technology
Irbid - Jordan
{natheer,hammouri, eyadtaq}@just.edu.jo

Abstract

In this paper we present the Collaborative Arabic Morphologic Service (CAMS). The main theme of this project is to provide the research community of Arabic morphology with a web based service that enables them to run various algorithms of without the need to program them on their own or migrate to different platforms. This has stemmed from the fact that, even though many algorithms in the field of Arabic natural language processing were presented, the majority of them do not provide the actual implementation for future use. Moreover, in the cases when such implementations are present, the users face the restriction of the lack of portability between different platforms. We present in this paper an intended architecture of this web service. Moreover, we discuss the internal specifics of communication between the users' end-systems and the server that hosts the service. In addition, we present a novel approach of merging the results of various algorithms instead of just resorting to one algorithm at a time. This, as a result, is expected to enhance the overall accuracy of the system. Finally, a logging and feedback unit will be incorporated such that the results of previous queries may be used in later similar queries.

Introduction

In the past few years, the research on computerizing the Arabic language has seen advances in multi directions. Several topics in Arabic natural language processing started to appear in the literature, such as concept based extraction, search engines, text-to-speech (synthesis), grammatical checking, and information retrieval. There is a desperate need for Arabic Morphological modules in such systems. Several algorithms for Arabic Morphological processing were published in the past decade. The majority of these algorithms did not provide the binary or the source code of their implementation (Imad et al, 2004). An example of an algorithm without an implementation is that in (Aljlayl & Frieder, 2002) which presented a linguistic based morphological analyzer with 87.4% increase in precision using the TREC benchmark data. Moreover, when the implementations of algorithms were made available, they were restricted to certain platforms which limited the accessibility of the users to them. For instance, (Al-Kharshai & Al-Sughayiyer, 2002) developed a Rule Based algorithm which was later implemented using C++. However, with such a case platform independence may not be guaranteed.

This, as a result, makes it hard for researchers to use these algorithms in further research related to Arabic language processing since they should implement them on their own to suit the platforms that they use. This issue raises the need for an Arabic Morphological module that is publicly available and easily accessible from different platforms.

In this paper we present the Collaborative Arabic Morphological Service (CAMS) which will implement various algorithms for Arabic morphological analysis. The goal is to provide a publicly available service without the constraints of cross platform usage.

System Architecture

As shown in Figure 1, each algorithm or technique is running on a separate PC. All PCs are connected to the server (central PC) via TCP connection. The server accepts the incoming requests and forwards them to all other PCs, which implement different Morphological algorithms. The server takes the result of each implemented algorithm and determines the most appropriate result using accuracy based voting mechanism. Moreover, the logging and feedback unit may be used to update the weights assigned to each voting entity.

Morphological Analyzer Algorithms

The first morphological analyzer is based on the Khoja's analyzer (Khoja & Shereen, 1999). This analyzer depends on several linguist database files such as a database that contains all siacritic characters, punctuation characters, definite articles, and a list of Arabic stop words referred by (Larkey & Connell, 2001). The analyzer starts by removing the longest suffix and the longest prefix of the given word, then it matches the remaining word by the given database files.

The second algorithm is based on the morphological analyzer presented in (Al-Shalabi et al, 2003). Unlike Khoja's analyzer which depends on linguistic database, this analyzer depends on assigning weights for each letter in the given word multiplied by the letter's position. Letters in the word "سالتمونيها" where all affixes and consonants were assigned a weight of zero. Letters with lowest weights were selected as root letters.

The third algorithm is based on Augmented Transition Network (ATN) (Saad et al). The search in ATN is improved by using some semantic rules of the Arabic language.

Accuracy Based Voting Mechanism

The results from different Morphological analyzers are fed to an accuracy based voting mechanism which computes a weighted sum of all the returned results. The weights are assigned relative to the accuracy of each individual algorithm. Thus, algorithms with high accuracy

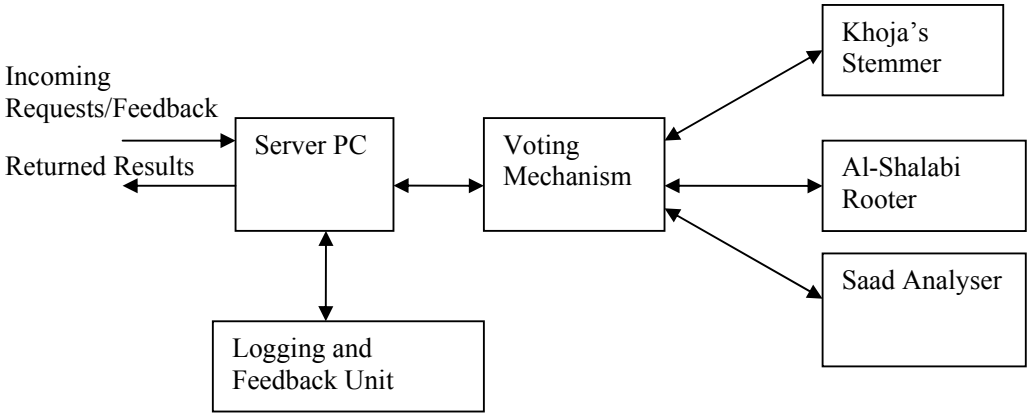

Figure 1: Overall architecture of the Collaborative Arabic Morphological Service

will contribute more towards the final decision compared to those less accurate.

Web Service Architecture

Figure 2 shows the web service architecture on which the service will be running on. The communication between the server and the clients requesting the service is handled using Simple Object Access Protocol (SOAP). This protocol is based on eXtensible Markup Language (XML) and Hypertext Transfer Protocol (HTTP). Data exchange is done using XML which is a general data format that can be integrated in different languages. The use of HTTP allows easier communication through proxies and firewalls (Alonso et al, 2004).

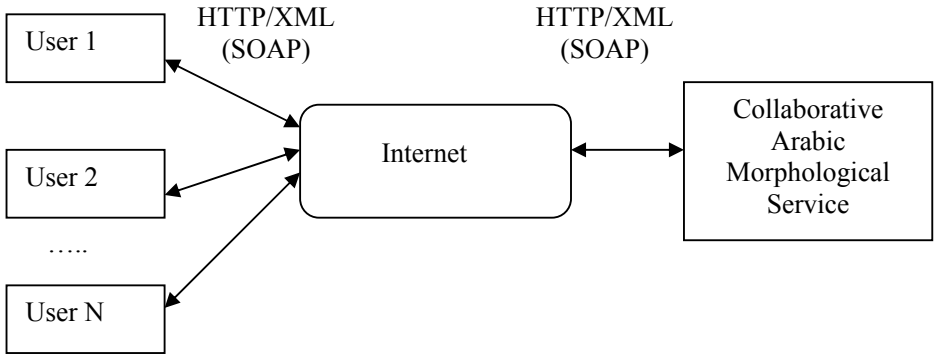

Figure 2: Web Service Architecture

System Interface

The service is implemented in C#.NET (.NET 2.0) and the data is stored in Microsoft SQL Server 2005. The service is hosted at Jordan University of Science and Technology computer center and can be accessed locally for testing. Service will be available soon for public at no charge for research purposes. The system usage is described through Web Services Description Languages (WSDL). WSDL is an XML document used to describe the Web services in terms of the location of the service and operations or methods the services provide. WSDL consists of abstract part and concrete part. Figure 3 shows the abstract part of the WSDL. The message element defines the name and type message. Port type defines the operations which are made up of messages.

Figure 4 shows the concrete part of the service. The binding part defines the name of the binding which can be any name and type points to the port type defined in the abstract part. The soap:binding attributes defines the style which can be rpc or document. The action is defined as transport protocol which is based on HTTP. The operation element defines each operation that the port exposes. The service part defines location of the service.

```
<wsdl:message name="GetRootSoapIn">
  <wsdl:part name="parameters" element="tns:GetRoot" />
</wsdl:message>
<wsdl:message name="GetRootSoapOut">
  <wsdl:part name="parameters" element="tns:GetRootResponse" />
</wsdl:message>

<wsdl:portType name="MorphologicalAnalyzerSoap">
  <wsdl:operation name="GetRoot">
    <wsdl:input message="tns:GetRootSoapIn" />
    <wsdl:output message="tns:GetRootSoapOut" />
  </wsdl:operation>
</wsdl:portType>
```

Figure 3: Abstract Part of the WDSL

```
<wsdl:binding name="MorphologicalAnalyzerSoap" type="tns:MorphologicalAnalyzerSoap">
  <soap:binding transport="http://schemas.xmlsoap.org/soap/http" />
  <wsdl:operation name="GetRoot">
    <soap:operation soapAction="http://arabic.just.edu.jo/GetRoot" style="document" />
    <wsdl:input>
      <soap:body use="literal" />
    </wsdl:input>
    <wsdl:output>
      <soap:body use="literal" />
    </wsdl:output>
  </wsdl:operation>
</wsdl:binding>


<wsdl:service name="MorphologicalAnalyzer">
  <wsdl:port name="MorphologicalAnalyzerSoap" binding="tns:MorphologicalAnalyzerSoap">
    <soap:address location="http://localhost:1252/ArabicService/Service.asmx" />
  </wsdl:port>
</wsdl:service>
```

Figure 4: Concrete Part of the WDSL

Summary and Future work

In this paper we presented the architecture of Arabic Morphological Analyzer and the underlying algorithms to be used. Moreover, an overview of the Web Service model to be implemented was discussed. Our next step is to build the actual system based on the architecture given in this paper. Once that is accomplished, we intend to fine tune the Accuracy Based Mechanism to produce the highest possible accuracy. This will be the first step in our continuous work towards building an Arabic sentence level syntactical service which breaks the Arabic sentence into: Verb, subject, object, adverbs, etc.

References

Alonso G., Casati F., Kuno H., and Machiraju V. Web Services Concepts, Architectures and Applications. Springer Verlag 2004

Castor, A. & Pollux, L.E. (1992). The use of user modelling to guide inference and learning. Applied Intelligence, 2(1), 37--53.

Chercheur, J.L. (1994). Case-Based Reasoning. San Mateo, CA: Morgan Kaufman Publishers.

Grandchercheur, L.B. (1983). Vers une modélisation cognitive de l'être et du néant. In S.G Paris, G.M. Olson, & H.W. Stevenson (Eds.), Fondement des Sciences Cognitives (pp. 6--38). Hillsdale, NJ: Lawrence Erlbaum Associates.

Martin, L.E. (1990). Knowledge Extraction. In Proceedings of the Twelfth Annual Conference of the Cognitive Science Society (pp. 252--262). Hillsdale, NJ: Lawrence Erlbaum Associates.

Imad A. Al-Sughaiyer, and Ibrahim A. Al-Kharashi, "Arabic morphological analysis techniques: a comprehensive survey", Journal of the American Society for Information Science and Technology, Vol.55 No.3, p.189-213, February 2004.

Zavatta, A. (1992). Un Générateur d'Insultes s'intégrant dans un Système de Dialogue Homme-Machine. Thèse de Doctorat en Informatique. Université Paris-sud, Centre d'Orsay.