Computers in Human Behavior 41 (2014) 92–103

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Word sense disambiguation using evolutionary algorithms – Application to Arabic language

Mohamed El Bachir Menai [*]

Department of Computer Science, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia

ARTICLE INFO

Article history:
Available online 3 October 2014

Keywords:
Natural language processing
Word sense disambiguation
Modern Standard Arabic
Evolutionary algorithms
Genetic algorithms
Memetic algorithms

ABSTRACT

Natural language processing is related to human–computer interaction, where several challenges involve natural language understanding. Word sense disambiguation problem consists in the computational assignment of a meaning to a word according to a particular context in which it occurs. Many natural language processing applications, such as machine translation, information retrieval, and information extraction, require this task which occurs at the semantic level. Evolutionary computation approaches can be effective to solve this problem since they have been successfully used for many real-world optimization problems. In this paper, we propose to solve the word sense disambiguation problem using genetic and memetic algorithms, and apply them to Modern Standard Arabic. We demonstrate the performance of several models of our algorithms by carrying out experiments on a large Arabic corpus, and comparing them against a naïve Bayes classifier. Experimental results show that genetic algorithms can achieve more precise prediction than memetic algorithms and naïve Bayes classifier, attaining 79%.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

A computer that could interact directly with human through a natural language interface has to deal with ambiguity. It is a key feature of natural languages. That is, words can have different meanings (polysemy), depending on the context in which they occur. Resolving ambiguity is one of the most difficult tasks in natural language processing. For example, in the three following sentences *"Go to school every day."*, *"The school has a blue façade."*, and *"The school is on strike."*, the word *"school"* means institution, building, and teacher, respectively. Humans deal with language ambiguities by acquiring and enriching common sense knowledge during their lives. However, solving computationally the ambiguity of words is a challenging task, since it relies on knowledge, its representation, extraction, and analysis. In Arabic language, ambiguity is present at many levels (Farghaly & Shaalan, 2009), such as homograph, internal word structure, syntactic, semantic, constituent boundary, and anaphoric ambiguity. The average number of ambiguities for a token in Modern Standard Arabic (MSA) is 19.2, whereas it is 2.3 in most languages (Farghaly & Shaalan, 2009). Arabic is a highly structured and derivational language where morphology plays a very important role (Farghaly & Shaalan, 2009; Attia, 2008; Beesley, 2001; Buckwalter, 2004b).

Word sense disambiguation (WSD) is a challenging task in the area of natural language processing (NLP). It refers to the task that automatically assigns the appropriate sense, selected from a set of pre-defined senses for a polysemous word, according to a particular context. Indeed, the identification of one word sense is related to the identification of neighboring word senses. WSD is necessary for many NLP applications and is believed to be helpful in improving their performance such as machine translation, information retrieval, information extraction, part of speech tagging, and text categorization. WSD has been described as an AI-complete (Artificial Intelligence-complete) problem (Mallery, 1988) that is analogous to NP-complete problems in complexity theory. It can be formulated as a search problem and solved approximately by exploring the solution search space using heuristic and meta-heuristic algorithms. Several approaches have been investigated for WSD in occidental languages (English, French, German, etc.), including knowledge-based approaches and machine learning-based approaches. However, research on WSD in Arabic language is relatively limited.

Evolutionary algorithms (EAs) are search and optimization methods inspired by biological evolution: natural selection and survival of the fittest in the biological world. Several types of EAs were developed, including genetic algorithms (GAs) (Holland, 1975), evolutionary programming (EP) (Fogel, Owens, & Walsh, 1966; Fogel, 1994), evolution strategies (ES) (Rechenberg, 1973; Rechenberg, 1994; Schwefel, 1965; Schwefel, 1993) and genetic

[*] Tel.: +966 11 4670687; fax: +966 11 4675423.
E-mail address: menai@ksu.edu.sa

http://dx.doi.org/10.1016/j.chb.2014.06.021
0747-5632/© 2014 Elsevier Ltd. All rights reserved.

programming (GP) (Koza, 1992). Memetic algorithms (MAs) (Moscato, 1989) are a combination of EAs with local search (also named hybrid GAs or genetic local search). EAs are among the most popular and robust optimization methods used to solve hard optimization and machine learning problems. They have been widely and successfully applied in several real world applications (Michalewicz, 1994) and research domains. These include NLP research, such as query translation (Davis & Dunning, 1996), inference of context free grammars (Keller & Lutz, 1997), tagging (Araujo, 2002), parsing (Araujo, 2001), and WSD (Gelbukh, Sidorov, & Han, 2003; Decadt, Hoste, Daelemans, & den Bosch, 2004; Zhang, Zhou, & Martin, 2008). Araujo (Araujo, 2007) has written a survey paper on how EAs are applied to statistical NLP, which is highly recommended.

In this paper, we study the potential of GAs and MAs in formulating and solving the WSD problem, apply them to MSA, and compare them with some existing methods. We implemented a system in which we experimented different variants of GAs and MAs for WSD. Our study results in the introduction of a competitive approach for WSD. The rest of the paper is organized as follows. The next section presents a brief overview of EAs. Section 3 introduces the WSD problem, and presents the main approaches to solve it. Section 4 describes Arabic language peculiarities and challenges. Section 5 presents the proposed approach to WSD and describes in detail the proposed algorithms. Section 6 reports and discusses the experimental results, and presents comparisons between the proposed algorithms and other methods. Finally, Section 7 concludes this paper, and outlines some future research directions.

2. Evolutionary algorithms

EAs are built around four key concepts (De Jong, 2006): population(s) of individuals competing for limited resources, dynamic changing populations, suitability of an individual to reproduce and survive, and variational inheritance through variation operators. EAs are categorized as "generate and test" algorithms that involve growth or development in a population of chromosomes in genotype space (individuals containing genes) of candidate solutions in phenotype space (real features of an individual). An evaluation function called the *fitness function*, defined from chromosome representation, measures how effective the candidate solutions are as a solution to the problem. Variation operators such as *recombination* (or *crossover* in case of recombination of two parents) and *mutation* are applied to modify the individual content and promote diversity.

Algorithm 1. Evolutionary algorithm

```
Initialize $P(1)$;
$t \leftarrow 1$;
while $\textit{not exit criterion}$ do
    evaluate $P(t)$;
    selection;
    recombination;
    mutation;
    survive;
    $t \leftarrow t + 1$;
```

The basic steps of an EA are outlined in Algorithm 1. An *initial population*, $P(1)$, is randomly generated and a selection process (*selection*) is then performed to select parents based on the fitness of individuals (*evaluate*). The *recombination* and *mutation* operators are applied on parents to obtain a population of offspring. The population is renewed (*survive*) by selecting individuals from the current population and offspring for next generation $(t + 1)$. This evolutionary process continues until a termination condition, *exit criterion*, is reached.

Algorithm 2. Genetic algorithm

```
input  : $Population_{size}$, $Problem_{size}$, $P_{crossover}$, $P_{mutation}$
output: $S_{best}$
$Population \leftarrow$ Initialize($Population_{size}$, $Problem_{size}$);
Evaluate($Population$);
$S_{best} \leftarrow$ BestSolution($Population$);
while $\textit{not exit criterion}$ do
    $Parents \leftarrow$ SelectParents ($Population$);
    $Children \leftarrow \phi$;
    for $Parent_1, Parent_2 \in Parents$ do
        $(Child_1, Child_2) \leftarrow$ Crossover ($Parent_1, Parent_2, P_{crossover}$);
        $Children \leftarrow$ Mutate ($Child_1, P_{mutation}$);
        $Children \leftarrow$ Mutate ($Child_2, P_{mutation}$);
    Evaluate($Children$);
    $S_{best} \leftarrow$ BestSolution($Children$);
    $Population \leftarrow$ SelectToSurvive ($Population, Children$);
Return($S_{best}$)
```

GAs (Holland, 1975) are the most traditional EAs which are based on biological genetics, natural selection, and emergent adaptive behavior. They are associated to the use of binary, integer, or real valued vectors for the chromosome representation. The crossover and mutation are the genetic operators. The crossover is the main operator (applied with a high probability), and the mutation is the secondary one (applied with a low probability). The main steps of a GA are outlined in Algorithm 2 (Brownlee, 2011). GP (Koza, 1992) can be considered as an extension of GAs in which each individual is a computer program represented by a rooted tree. In this case, the fitness function determines how well a program is able to solve the problem. MAs (Moscato, 1989) (also called genetic local search algorithms) are considered as hybrid EAs which use one or more local search phases within GAs. While GAs rely on the concept of biological evolution of a population, MAs combine the individual learning with the evolutionary adaptation of a population (cultural evolution). Algorithm 3 describes a standard MA. Note that a local search (Algorithm 4) is introduced within a GA to mimic the individual learning of population members by searching a local neighborhood to identify a better solution.

Algorithm 3. Memetic algorithm

```
input  : $Population_{size}$, $Problem_{size}$, $P_{crossover}$, $P_{mutation}$
output: $S_{best}$
$Population \leftarrow$ Initialize($Population_{size}$, $Problem_{size}$);
for $Individual \in Population$ do
    $Individual \leftarrow$ Local-Search($Individual$);
Evaluate($Population$);
$S_{best} \leftarrow$ BestSolution($Population$);
while $\textit{not exit criterion}$ do
    $Parents \leftarrow$ SelectParents ($Population$);
    $Children \leftarrow \phi$;
    for $Parent_1, Parent_2 \in Parents$ do
        $(Child_1, Child_2) \leftarrow$ Crossover ($Parent_1, Parent_2, P_{crossover}$);
        $Child_1 \leftarrow$ Local-Search($Child_1$);
        $Child_2 \leftarrow$ Local-Search($Child_2$);
        $Child_1 \leftarrow$ Mutate ($Child_1, P_{mutation}$);
        $Child_2 \leftarrow$ Mutate ($Child_2, P_{mutation}$);
        $Children \leftarrow$ Local-Search($Child_1$);
        $Children \leftarrow$ Local-Search($Child_2$);
    Evaluate($Children$);
    $S_{best} \leftarrow$ BestSolution($Children$);
    $Population \leftarrow$ SelectToSurvive ($Population, Children$);
Return($S_{best}$)
```

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات