Towards Resolving Morphological Ambiguity in Arabic Intelligent Language Tutoring Framework

Khaled Shaalan[1], Marwa Magdy[2], Doaa Samy[3]

[1] The British University in Dubai, PO Box 345015 Dubai, UAE

[2] Faculty of Computers & Information, Cairo University, 5 Ahmed Zewel St., Giza 12613 Egypt

[3] Cairo University

khaled.shaalan@buid.ac.ae, m.magdy@fci-cu.edu.eg, doaasamy@cu.edu.eg

Abstract

Ambiguity is a major issue in any NLP application that occurs when multiple interpretations of the same language phenomenon are produced. Given the complexity of the Arabic morphological system, it is difficult to determine what the intended meaning of the writer is. Moreover, Intelligent Language Tutoring Systems which need to analyze erroneous learner answers, generally, introduce techniques, such as constraints relaxation, that would produce more interpretations than systems designed for processing well-formed input. This paper addresses issues related to the morphological disambiguation of corrected interpretations of erroneous Arabic verbs that were written by beginner to intermediate Second Language Learners. The morphological disambiguation has been developed and effectively evaluated using real test data. It achieved satisfactory results in terms of the recall rate.

1. Introduction

An Intelligent Language Tutoring System (ILTS) is a computer-based educational system that allows simulation of a human tutor. An ILTS is a valuable tool used in language e-learning programs. Besides, it is highly demanded as an application within the Natural Language Processing field since it helps people in the language learning process either for native or for foreign languages. These NLP tools used in language learning can be used in several ways such as *parsing* of the learner input and *diagnosis* of morphological and syntactic errors (Nerbonne, 2003). However, ILTS for error diagnosis to analyze learners' input and provide intelligent and real-time feedback is highly needed for the following reasons:

- ILTS provide individualized tutoring to learners who are often left to themselves and cannot rely upon teachers and tutors to help them.
- Reliable error diagnosis systems would allow users/authors to overcome the limitations of multiple choice questions and fill-in-the-blanks types of exercises. Besides, ILT systems can provide a suitable platform for introducing more communicative and interactive tasks to learners (L'haire and Faltin, 2003).

Unfortunately, almost all NLP tools such as parsers, morphological analyzer, etc, are designed to handle well-formed input. So, to handle ill-formed input in ILTS, techniques such as constraint relaxation are employed (Faltin, 2003). In any language model, the partial structures can combine only if some constraints or conditions are met. When these constraints are relaxed, an attachment is allowed even if the constraint is not satisfied. The relaxed constraint must be marked on the structure such that the type and position of the detected error can be indicated (confirmed) later on. In ILTS, relaxing the constraints of the language to analyze learner's answer inevitably produce ambiguous solutions, i.e., more corrected interpretations, than systems designed for only well-formed input (Attia, 2006). Consider, for example, the learner input Arabic word عيشت. This would have two interpretations: 1) the learner might mean عشت /Ei$tu/[1] (lived-I) which is related to problems with vowel letters that makes the short vowel الكسرة /i/ long one ياء /y/, or 2) s/he might mean عيشت /Eay~a$tu/ (sustained-I).

This paper addresses issues related to the morphological disambiguation of corrected interpretations of erroneous Arabic verbs written by beginner to intermediate Second Language Learners (SLLs). The proposed system follows the approach a language teacher uses in disambiguating and selecting a preferred analysis. It considers the likelihood of an error which takes into account the level of instruction and the frequency and/or difficulty of Arabic concepts. The concern here is to avoid misleading or incorrect feedback. The result of disambiguation and selecting appropriate analysis is used within ILTS framework to detect the exact source of error and provide the error specific feedback.

Ahmed (2000) addressed the problem of Arabic morphological disambiguation to select the most likely morphological analysis for each well-formed word in the text. He used a powerful dynamic n-gram statistical disambiguation technique. The statistical knowledge of the system may be altered or adjusted anytime to consider any desired text corpus. But, to the best of our knowledge no research has addressed the problem of disambiguating *corrected* interpretations of ill-formed Arabic verbs.

The rest of this paper is structured as follows. Section 2 presents a brief discussion of Arabic morphological ambiguity problem. Section 3 describes the proposed system. Section 4 discusses the results from the conducted experiment. Finally, in Section 5, we give some concluding remarks.

[1] Buckwalter transliteration is used here to Romanize Arabic examples (Buckwalter 2002).

2. Arabic Morphological Ambiguity Problem

Arabic language is one of the Semitic languages that is defined as a *diacritized* language where the pronunciation of its words cannot be fully determined by their spelling characters only. Diacritics are special marks put above or below the spelling characters to determine the correct vocalization and, thus, the correct pronunciation.

Unfortunately, diacritics are rarely used in current Arabic writing conventions. The correct pronunciation and interpretation of none or partially diacritized text depends on the native language competence and the context. Due to the optional diacritization, two or more words in Arabic are homographic: they have the same orthographic form, though the pronunciation and meaning is totally different (Ahmed, 2000; Attia, 2006; Habash, 2004). Table 1 listed some homographic examples.

Word	Lemma	Different Interpretations
يعد /yEd/	أعاد />aEAd/	يعِد /yuEid/ (bring back)
	عاد /EAd/	يعُد /yaEud/ (return)
	وعد /waEid/	يعِد /yaEid/ (promise)
	عد /Ead~/	يَعُدّ /yaEud~/ (count)
	أعد />aEd~/	يُعِدّ /yuEid~/ (prepare)

Table 1: An Arabic word that is homographic

However, other factors contribute to the problem of morphological ambiguity in Arabic. Among these factors (Attia, 2006):

1. Orthographic alteration operations (such as deletion) frequently produce inflected forms that can belong to two or more different lemmas as shown in Table 1. These alteration operations are due to the phonological constraints of certain root consonants. The important irregularity issues are related to Arabic weak verbs that include one or more weak letter. Weak letters can be deleted or substituted by other letters because of Arabic phonological constraints (El-Sadany and Hashish 1989). For example, the deletion of the letter (و) in taking the present (imperfect) tense of the trilateral root و-ع-د /w-E-d/, using regular rules would generate يوعد* /ya-wEid/ but as it is a assimilated (first weak) verb it should be generated according to special weak rules and thus it appears in written texts as يعد /ya-Eid/ (promise).
2. Some Arabic patterns are different only in that one of them has a doubled sound which is not explicit in writing of their corresponding forms such as فعل /faEala/ and فعّل /faE~ala/.
3. Many inflectional operations underlie a slight change in pronunciation without any explicit orthographical effect due to lack of short vowels (diacritics). An example of this is the ambiguity of active vs. passive vs. imperative verb forms.
4. Some prefixes and suffixes can be homographic with each other. For example, the perfect verb suffix ت /Teh/ can indicate either: 1) first person singular, 2) second person singular masculine, 3) second person singular feminine, or 4) third person singular feminine.
5. Prefixes and suffixes can accidentally produce a form that is homographic with another full form word. For example, the word أسد can be interpreted as أسد />asad/ (lion) or أسَّدَ />a-sud~/ (I-Block).

Difficulties in the process of Arabic morphological disambiguation are the main reason behind addressing the challenges of developing a morphological disambiguation module/tool/ etc that can handle ill-formed Arabic verbs.

3. The Proposed Disambiguation System

The proposed system is an integral part of an Arabic ILTS for SLLs. The system is cable of analyzing both well- and ill-formed learner answers. The ILTS analyzes each input word and produces all of its possible analyses (Shaalan, Magdy and Fahmy, 2010). Afterwards, the ILTS sends these analyses to the disambiguation system to select the appropriate analysis. The selected analysis is then used to detect the exact source of error introduced by the learner and, consequently, the ILTS generates a full diagnosis of the learner input. This is clarified by the following figure.

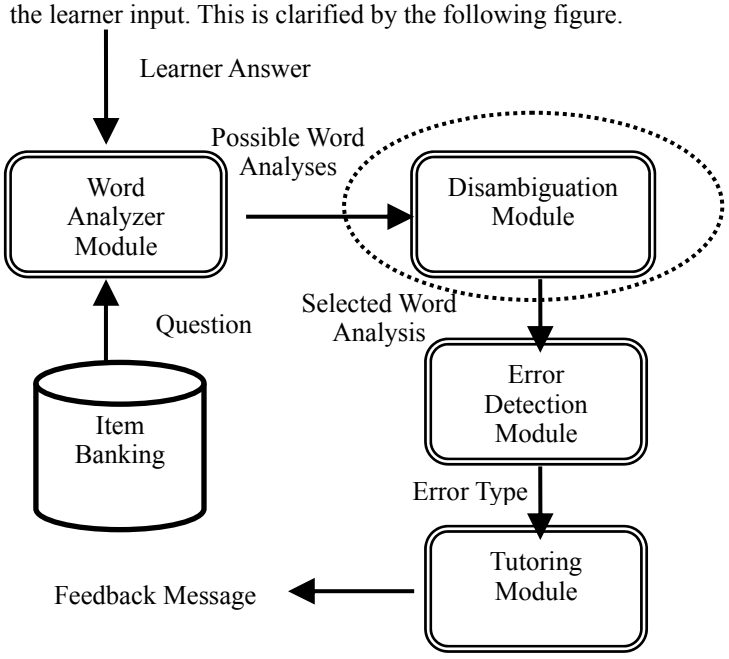

Figure 1: Arabic ILTS Framework

The following example clarifies how the system works. Consider the following question that is presented to the

learner:

Example 1:
Complete the following sentence with the correct conjugation of the given root in imperfect tense active voice.

..... (ب-ي-ع) جدتي الارز
/…. (b-y-E) jad~atiy Al>aruz~/ [my grandmother .... (sell) the rice]

In the above example, the root ب-ي-ع /b-y-E/ contains middle weak letter ي /y/ so it needs special rules to conjugate it in different forms. For example to conjugate it into imperfect passive voice, the middle weak letter should be substituted by ا /A/ so it become تُباع /tu-baAE/ (was sold)

Assume the following two answers; where (a) includes a wrong conjugation of a *Hollow* (middle weak) verb, and (b) is the correct answer.

a. تباع جدتي الارز /ta-biAE jad~atiy Al>aruz~/ (My-grandmother sells the-rice).
b. تبيع جدتي الارز /ta-biyE jad~atiy Al>aruz~/ (My-grandmother sells the-rice).

The ILTS produces two possible analyses for the erroneous word تباع:

- *Third person singular feminine imperfect verb in the active voice with converted middle letter ي /y/ to ا/A/.*
- *Third person singular feminine imperfect verb in the passive voice.*

Then the disambiguation system selects the most appropriate analysis according to: the learner level and difficulty of Arabic concepts[2]. For example in Arabic, the passive voice is a rare construction and it is doubtful that a beginner learner of Arabic would write a passive voice of a verb instead of its active voice. Therefore, the system adopts some *prioritized conditions* to select the most preferred word analysis. Hence, in this case, the system will select the *first analysis*. This analysis is later on used by ILTS to detect the error made by the (incorrect conjugation of verb in imperfect tense active voice)

In the proposed system, we investigated our disambiguation approach on the following three types of ambiguous analysis of erroneous learner input:

1. The orthographic match in non-diacirtized text between Arabic conjugated verb forms in passive voice, and active voice, imperfect or perfect tense, respectively. For example, (نَقَل */naqala/*) is the perfect tense of the 3rd person singular masculine in active voice, while (نُقِل */nuqil/*) is the perfect tense for the 3rd person singular masculine in passive voice. Same phenomenon is repeated in the imperfect tense (يَنقُل|يُنقَل */yanqul|yunqal/*)
2. The orthographic match between different affixes in terms of spelling characters. These affixes are used to conjugate different verb forms. For example the prefix (ت) can be used to conjugate the present tense of the 3rd person feminine singular (هي تذهب) and the 2nd person masculine singular (أنت تذهب)
3. The orthographic match between Arabic verb derivation patterns and non-derivative patterns. For example, the verb سعد /saEada/ (to be happy) is a root, non-derivative verb. A possible derivative pattern is أسعد/AsEada/(to make happy). The imperfect conjugation for the first person of the first verb is (أسعد */AsEada/*), which is identical to the conjugation of the 3rd person singular in the perfect tense of the second verb (هو أسعد */AsEada/*).

There are some other types of ambiguities[3] that are out of the scope of the current system as the system has no direct knowledge of what the student meant to express. In some systems, where the system has insufficient knowledge to proceed with, a dialogue is established with the learner in order to guide the selection of appropriate expression, e.g. (Hsieh et al., 2002). Figure 2 presents how the system disambiguates multiple analyses and the rest of this section explains in more details.

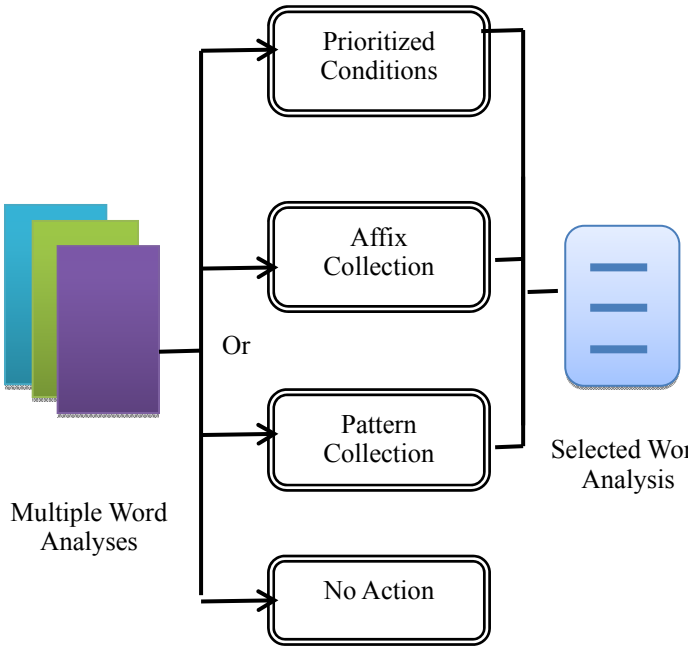

Figure 2: Disambiguation System Structure

In case of the *first ambiguity type*, the system selects the word analysis a student most likely intended. It implements *two* prioritized conditions to selects the most preferred word analysis:

1. If the question goal is to test *passive voice* then the system selects *passive voice* analysis; otherwise, it selects the *active voice* analysis, or

[2] This rule is applied by Arabic language teacher (Heift, 1998).

[3] Example of these types is when the *noun* has the same orthographic form as *verb*

2. If the question goal is to test *imperative tense* then the system selects the *imperative tense* analysis; otherwise, it selects the *perfect or imperfect tense* analysis.

By this way, in Example 1, the system applies the first condition to select the first analysis (*Third person singular feminine imperfect verb in the active voice*). Notice, however, that the question objective is to test conjugation of imperfect active voice verb.

In case of the *second ambiguity type* (i.e. orthographic match between different affixes), the system collects all affixes with the same orthographic form but which differs in their morpho-syntactic features in one entry with a generic feature structure.

For example, consider the following learner input; where (b) is the correct answer:

a. محمد تورطت في جريمة قتل /muHam~ad tawar~aTt fiy jariymap qatol/ (Mohamed was-involved in murder crime).

b. محمد تورط في جريمة قتل /muHam~ad ta-war~aTa fiy jariymap qatol/ (Mohamed was-involved in murder crime).

The learner here has made a subject-verb disagreement between the subject Mohamedمحمد and the verb was-involved تورطت. Four possible analyses of the erroneous verb are produced:

- *First person singular perfect verb in the active voice.*
- *Second person singular masculine perfect verb in the active voice.*
- *Second person singular feminine perfect verb in the active voice.*
- *Third person singular feminine perfect verb in the active voice*

These four possible analyses are combined into the generic analysis:

- *Singular perfect verb in the active voice.*

In case of the *third ambiguity type* (i.e. orthographic match between different patterns), the system collects all these patterns in one entry with a generic feature structure.

For example, consider the following question that is presented to the learner:

Example 2:
Complete the following sentence with the correct conjugation of the given root in perfect tense active voice.

جدي وجدتي .... (ن-ق-ل) إلي بيت جديد
/jad~iy wajad~apiy …. (n-q-l) <ilaY bayot jadiyd/ (my grandfather and my grandmother .... to a new house)
Assume the following learner input; where input (b) is the correct answer:

a. جدي و جدتي نقلوا إلي بيت جديد / jad~iy wajad~apiy naq~aluwA <ilaY bayot jadiyd / (my-grandfather and my-grandmother moved to a new house).

b. جدي وجدتي انتقلا إلي بيت جديد /jad~iy wajad~apiy {inotaqalA <ilaY bayot jadiyd/ (my-grandfather and my-grandmother moved to a new house).

The learner here has made two errors: 1) subject-verb disagreement between the subject "my-grandmother and my-grandfather جدي وجدتي" and the verb "نقلوا", the subject is dual while the verb is conjugated in the masculine plural form and, 2) incorrect use of the root pattern of a perfect verb form; the correct pattern is 'افتعل' while the learner used the pattern 'فعل'. However, the ILTS produced two possible analyses as shown in the following:

- *Third person masculine plural perfect verb in the active voice following the pattern 'فعل'.*
- *Third person masculine plural perfect verb in the active voice following the pattern 'فعّل'.*

These two possible analyses are combined into generic feature structure:

- *Third person masculine plural perfect verb in the active voice.*

4. Experiment

We conducted an experiment that measures how successfully the proposed model selects the most appropriate analysis that is used later on to detect the exact source of error the learner has made. The *quantitative* measures are used. These measures rely on collecting different test sets written by real SLLs in a typical teaching/learning environment. It was necessary that these learners have different backgrounds (i.e., differ in their first language) to test if the system is general enough and not aimed to a specific sort of learners. The test set is then fed into the system and the solved ambiguous cases and unsolved are reported. The recall rate is calculated. This measure has been used in evaluating similar research (cf. Wagner et al., 2007; Sjöbergh and Knutsson 2005; Faltin 2003).

The abovementioned methodology is applied on a real test set that consists of 116 real Arabic sentences. The number of words per sentence varies from 3 to 15 words, with an average of 5.1 words per test sentence. The total number of words in all test sentences are 587 words, 118 of them have lexical verb errors. 72 verbs are ambiguous cases. The system successfully solved 46 cases of them while it failed to select the correct analysis for 26 cases. The next section will discuss all failed cases.

4.1 Evaluation Problems Classification

In this section, we discuss all problems which the proposed system failed to select the correct analysis. The major problem is it is difficult to determine what the intended meaning of the learner given the complexity of Arabic language.

The 26 failed cases are classified as follows:

- *Orthographic match between un-vocalized forms*. Arabic ILTS handles un-vocalized rather than vocalized

written Arabic text. This leads sometimes to more than one possible match between the same and different word categories. The total number of occurrences of this category is 8 cases. They are classified as follows:

- *Orthographic/homographs match between verb and noun forms*. This case happens when an Arabic verb has the same orthographic form as a noun. For example, consider the word تناول; it can lead to three possible correct words. It is not clear whether the learner meant the word to be: 1) the noun تناول /tanAwul/ (dealing with/ eating), 2) the perfect verb تناول /tanAwala/ (he/it-dealt with/ ate), or 3) the imperfect verb تناول /tu-nAwil/ (hand over/ deliver). The total number of occurrences of this problem is 7 cases.
- *The special case of the orthographic match between the Arabic third person singular perfect verb following the pattern أفعل /*>*afoEal/ and the first person singular imperfect verb as the word أوقع/.* It can lead to two possible interpretations. It is not clear whether the learner meant the word to be: 1) the perfect verb أوقع />awoqaEa/ (he/it-inflicted), or 2) imperfect verb أوقع />u-waq~iE/ (I-sign). The total number of occurrences of this problem is one case.

- *Additional- orthographic matches as a result of relaxing a constraint.* Applying the constraints relaxation technique in order to be able to analyze erroneous learner answers sometimes introduces extra orthographic matches. The total number of occurrences of this category is 18 cases. They are classified as follows:
  - *Orthographic matches produced for Arabic verbs after relaxing the long vowel to the short one.* For instance, consider the erroneous word هجر. It is not clear whether the learner meant the word to be: 1) هاجر /hAjara/ (he/she/it-emigrated) by making the long vowel a short one, 2) هجَّر /haj~ara/ (he/it-deported) by using the pattern فعَّل /faE~al/, 3) هجر /hajara/ (he/it-left) by using the pattern فعل /faEal/, or 4) هجر /hajor/ (abandoning) by using nouns instead of verbs. The total number of occurrences of this problem is 8 cases
  - *Orthographic matches produced after allowing incorrect conjugation of a verb.* For instance, consider the erroneous word أجوب. It is not clear whether the learner meant the word to be: 1) the imperfect verb أجيب />u-jiyb/ (I-answer), 2) or imperfect verb أجوب />a-juwb/ (I-explore). The total number of occurrences of this problem is 7 cases
  - *Orthographic matches produced for Arabic verbs after relaxing the short vowel to the long one.* For instance, consider the erroneous word عيشت. It is not clear whether the learner meant the word to be: 1) عشت /Ei$-tu/ (I-lived) by making the short vowel a long one or, 2) عيشت /Eay~a$-tu/ (I-sustained) with using the pattern فعَّل /faE~al/. The total number of occurrences of this problem is 2 cases.
  - *Orthographic matches produced after allowing incompatible usage of connected pronouns.* For instance, consider the erroneous word أعملت. It is not clear whether the learner meant the word to be: 1) the perfect verb أعملت />aEomal-tu/ (I-employed) or, 2) the perfect verb عملت /Eamiltu/ (I-worked) by using incompatible pronouns أ, ت (Alef, Teh). The total number of occurrences of this problem is one case.

Notice, however, that we asked human linguists about failed cases and he has identified most of theses cases as ambiguous.

5. Conclusion

The ambiguity problem is a standard problem in any NLP application. It is the major reason why computers do not yet understand natural language. However, the ambiguity problem presents a challenge to ILTS. That is because selecting the wrong analysis of student input can lead to misleading feedback or an error might be overlooked. Beside that given the complexity of Arabic language, this makes the ambiguity a serious problem and needs to be resolved. The preferred method in ILTS for disambiguating multiple readings of a wrong answer should consider the likelihood of an error and the difficulty of concepts. But with the lack of erroneous corpus, we depend on some linguistic studies that investigate the likelihood of errors. However, the ambiguity problem cannot be resolved totally and there is a need to issue a dialogue with the learner to know what exactly he means. Moreover, if a large tagged erroneous corpus exist then the ambiguity problem can be resolved by considering the likelihood of errors

6. References

Ahmed, M. A. 2000. A Large-Scale Computational Processor of the Arabic Morphology, and Applications. Master thesis, Cairo University, Egypt.

Attia, M. A. 2006. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks. In Proceedings of the Challenge of Arabic for NLP/MT Conference, 2006. The British Computer Society, London.

Buckwalter, T. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania, LDC Catalog No.: LDC2002L49, ISBN 1-58563-257-0.

El-Sadany, T. A. and Hashish, M. A. 1989. An Arabic Morphological System. In IBM Systems Journal, 28(4): 600- 612.

Faltin, A. V. 2003. Syntactic Error Diagnosis in the Context of Computer Assisted Language Learning. PhD thesis, University of Geneva, Switzerland.

Habash, N. 2004. Large Scale Lexeme Based Arabic Morphological Generation. In Proceedings of Traitement Automatique du Langage Naturel (TALN-2004). Fez, Morocco.

Heift, T. 1998. Designed Intelligence: A Language Teacher Model. Ph.D. Thesis, Simon Fraser University, Canada.

Hsieh, C.-C., Tsai, T.-H., Wible, D. and Hsu, W.-L. 2002. Exploiting Knowledge Representation in an Intelligent Tutoring System for English Lexical Errors. In Proceedings of the International Conference on Computers in Education ICCE 2002, Auckland, New Zealand, pp: 115-116.

L'haire, S. and Faltin, A. V. 2003. Error Diagnosis in the FreeText Project. In Calico Journal, 20 (3): 481-495.

Nerbonne, J. 2003. Natural Language Processing in Computer-Assisted Language Learning. In Ruslan Mitkov, editors, the Oxford Handbook of Computational Linguistics. Oxford, pp: 670-698.

Shaalan, K., Magdy, M., and Fahmy, A. 2010. Morphological Analysis of Ill-formed Arabic Verbs in Intelligent Language Tutoring Framework. In Proceedings of FLAIRS-23, Daytona Beach, Florida, USA. To appear.

Sj¨obergh, J., and Knutsson, O. 2005. Faking Errors to Avoid Making Errors: Machine Learning for Error Detection in Writing. In Proceedings of RANLP 2005, Borovets, Bulgaria, pp: 506-512.

Wagner, J., Foster, J., and Genabith, J. V. 2007. A Comparative Evaluation of Deep and Shallow Approaches to the Automatic Detection of Common Grammatical Errors. In Proceedings of EMNLP-CoNLL 2007, Prague, Czeck Republic, pp: 112-121.