World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 6, No. 1, 8-11, 2016

Speculation and Negation Annotation for Arabic Biomedical Texts: BioArabic Corpus

Fatima T. AL-Khawaldeh
Department of Computer Science
Al-Albayt University
Al-Mafraq, Jordan

Abstract—Negation and speculation are two common linguistic concepts in natural language processing field, need more semantic understanding of texts. They are used to definite factuality of text. Negation is used to express the opposite of the text and the Speculation is used to determine the degree of certainty. Biomedical text mining is the main natural language processing application concerns with negation and speculation to distinguish between facts and uncertain or negated information in biomedical text. To our knowledge, there is no previous research on annotating Arabic biomedical text to identify the negative or speculative expression and no publicly available standard corpora of suitable size that are practical for evaluating the automatic detection of negation and speculation tools and scope determination. This paper presents produced corpus handling negation and speculative in Arabic biomedical texts with the main annotation (we call this corpus the BioArabic corpus). The goal of building BioArabic corpus is to help biologists and computational linguistics, who develop tools for identifying the negation and speculation, to train and evaluate these tools since in biomedical texts language, assumptions, experimental results and negative results are used extensively. We will report our statistics on corpus size and the consistency of annotations.

Keywords-Arabic NLP; negation; speculation; biomedical (medical and biological); cues; certainty.

I. INTRODUCTION

Wide variety of natural language processing tasks, including sentiment analysis, question answering , and medical data mining concern with detecting the negation and speculation. In medical data mining, the growth of biological literature number and the rapid increasing of biomedical research papers published by numerous of journals, academic sites and other publishers increased the need for language processing methods to distinguish between facts and uncertain or negated information. One of the main necessary tasks in most text mining tasks is recognizing negative and speculative information such as in sentiment analysis [1].

Computational linguists seek to develop tools to biologists, help them to get necessary statements (facts, speculative or negated information) suitable to their need and remove the needless information. Automatic detection of speculative and negated sentences is one of the most useful tools in biomedical text mining to distinguish uncertain information from factual information according to biologist need. Biologists may concern to get uncertain information rather factual information for special research needs.

Speculation is defined as the existing of claimed thing but not sure [2]. More restrictive meaning of the word speculation is appeared if the biologists care of obtaining all speculations of biological thing, it is meant by hypothesis and uncertain [3]. The main goal of speculation detection is extracting all sentences expressing uncertainty [4]. Speculations is considered to be relevant information to biologist [3]. Negation turns a positive statement into negative. In the example, (انها تدرسAnhAtdrs/ انها لا تدرسAnhAlAtdrs),لا lA is negative particle negated the verbتدرسtdrs.

In the medical domain, negation and speculation identification plays essential role to mark all possible analyses and provide information that compares with the helpfulanalysis in biomedical scientific articles and abstracts. [5]. In order to help improvement the training and evaluation for speculation and/or negation information extraction tools, some annotated corpora are freely available. Bioscope is one of the largest annotated corpus for speculation, negation and its linguistic scopes in biomedical texts consists of three parts: medical free texts, biological full papers and biological scientific abstracts texts [6]. The authors in [7] annotated biological events with negation and uncertainty in the GENIA event corpus.

[1]In order to provide more meaning to the identification of the negation and speculation, the scope of the negation and speculation is denoted. The scope is the grammatical part in a sentence that is negated by negative cue or speculated by speculation cue aims to determine the linguistic coverage of negative keywords or speculative keywords [8].

To our knowledge, there is no previous research on annotating Arabic biomedical texts. The contributions of this paper are the following:

- We present an annotated corpus freely available in order to be used by Computational linguists, biologists and researchers and to help them to improve the detection of speculative and negated systems. Corpora annotated for negation and speculation are essential for the training and testing.
- We put some guidelines for distinguishing between facts, speculative sentences and negative sentences in Arabic biomedical texts.

II. Related Work

Recently, the interest for identifying negative and speculative language in natural language processing tasks has grown but there is a limited amount of literature studies in this domain. Since the biomedical texts include more hypothesis, negated results and uncertain sentences, the main focus was on biomedical texts (biological and medical scientific articles and abstracts).

In 2004, the authors in [9], asked four annotators to annotate 891 biomedical scientific abstracts sentences to three level of speculation: highly speculative, low speculative and fact.11% of all sentences are speculative. They concluded that the end of abstracts has the most of speculative sentences and obtained good results kappa between 0.54 and 0.68.In 2008, in [2]bioscope corpusis produced from more than twenty thousand medical sentences and specified guidelines to three annotators for annotating the sentence of biological full paper and abstracts to speculative and negation keywords along with their scope. 14% of speculative keywords is got from the all sentences. F1-score for negation keywords was between 0.91 and 0.96, and F1-score for speculative keywords was between 0.84 and 0.92. In 2010, Swedish medicalcorpus was created and evaluated for negated and speculative keywords using a few basic guidelines and rules[10]. This corpus contains negation words, speculative words, uncertain expressions and certain expressions. For training and testing, they used bioscope corpus, obtained a precision of % 97.6 and a recall of % 96.7 for negation cues for English depending on bioscope corpus. The authors of [3] developed a rule-based system to annotate speculative previous and new sentences to extract types of speculations.They showed the efficiency of their BioExcom corpus experienced on bioscope corpus and proved its essential role of biologists.

In [11],the authors presented a survey to describe the role of negation in sentiment analysis such the negation effects on the polarity of opinion and showed that its necessary role. In 2015, a machine learning approach was presented to automatically detecting negation and speculation for Sentiment Analysis by identifying negation and speculation cues the determining the full scope of these cues. The authors showed that this system obtained better results than the baseline such in the negation cue detection the results was 20% and 13% in the scope recognition[1].

For Arabic language, in 2015, the author of [12] showed that better performance was obtained when detecting the entailment relation and non-entailment relation by resolving the negation.

III. BioArabic Corpus

The existing of Arabic biomedical corpus is essential resource to simplify the process of developing speculation and negation detection systems for Arabic biomedical texts. Another significant role of available Arabic biomedical corpus is to facilitate the training end evaluation of thesesystems.

TheBioArabic corpus consists of seventy medical and biological papers texts taken from three dissimilar publishers (Iraqi Journal of Biotechnology المجلة العراقية (للتقانات الحياتية),Journal of Damascus University for Health Sciences (مجلة جامعة دمشق للعلوم الصحيه), Biotechnology News (اخبار التقانه الحيويه)), in order to guarantee the reliability and consistency in biomedical domain.

The BioArabic annotation process was carried out by two main stages: the first stage is: the determination of negative and speculative cues and the second stage is: the identification of negation and speculation cues linguistic scope. Five linguist annotators were provided by the guidelines concluded by our linguist experience and from previous guidelines of bioscope corpus, shared to mark the data. When the five annotators completed the annotations process, if there were two annotators or more have similar annotation, this annotation is taken. If there was not common annotation, we depend on our linguistic expert to select the most suitable annotation. The annotation process specifies the borders of the keywords and their scope.

BioArabic corpus consists of 10165 sentences, 26.2% of these sentences have linguist annotation, including negation words and speculative words.

IV. Annotation Guidelines

We follow some guidelines to annotate the BioArabic corpus where the guidelines from 1-9 are similar to bioscope corpus guidelines [6], and the guidelines from 10-13 are concluded from our experimental analysis.

- Guideline 1: Only sentences have one or more speculative or negative words are annotated.

[1] Cues and keywords are used in this paper interchangeably which mean the speculation and negation words.

- Guideline 2: Sentences don’t include any negative or speculative cue are left and disregarded.
- Guideline 3: questions are not annotated and ignored since they are already ambiguity.
- Guideline 4: The sentences includes any of negative keywords are annotated as negated sentences.
- Guideline 5: The sentences includes any of speculative keywords are annotated as speculative sentences.
- Guideline 6: Scopes of speculative or negative keywords are denoted in their keywords (including their keywords), scope is denoted by parentheses.
- Guideline 7: Speculative keywords are denoted in angled brackets, for example: يمكنymkn,<يمكن>.
- Guideline 8: Negative keywords are denoted in square brackets for example: لم*lam* ,[لم].
- Guideline 9: scope for each part in a sentence that is negated or speculated, is denoted where the scope is something is negated or speculated. Scopes are extended to the largest grammatical unit including the keywords.
- Guideline 10: form Arabic linguistic expert, the scope of speculative keywords (verbs, adjectives) usually starts after the keyword in the case of the speculative keywords are (adverbs) it starts before the speculative keywords and may include the whole sentence.

- Guideline11: If the scope is omitted part, for simplicity, we put the scope is (انه)Anh.

- Guideline12:The most speculative keywords are: يبينybyn/يحتمل yHtml/يسأل ysOl/يظهر yZhr/ يعتقدyEtqd/يدل ydl/اما...أو AmA...Ow/يقترح yqtrH/ قدqd/يمكن ymkn/أن On/يستطيع ystTyE/يفترض yftrD/ يبدوybdw/يتوقع ytwqE/يخمن yxmn/نسبيا nsbyA/أن يكون On ykwn/يشيرy$yr/ينفيynfy, etc...

- Guideline13:The most negative keywords are the particle negative particles and others:*maa (ما)*, the particle *laa(لا)*, the particle *lam* (لم) ,the particle *lan*(لن), and the *particle laysa*(ليس)And others like: عدم Edm/غير gyr/ مستحيلmstHyl etc…

V. BIOARABIC CORPUS ANNOTATION EXAMPLES

Example 1:

كما (< يعتقد>ان طفرات Missenses الواقع في الجزء الخارجي من بروتين المستقبل) تؤثر على فعالية المستقبل من خلال تناقض الفة ارتباط الهرمون المحفز TSH بالبروتين المستقبل TSHR.

kmA yEtqd An TfrAt Missenses AlwAqE fy Aljz' AlxArjy mn brwtyn Almstqbl tWvr ElY fEAlyp Almstqbl mn xlAl tnAqD Alfp ArtbAT Alhrmwn AlmHfz TSH b Albrwtyn .Almstqbl TSHR

- يعتقدyEtqd is speculative word(cue) and denoted by angled brackets<>.
- ان طفرات Missenses الواقع في الجزء الخارجي من بروتين المستقبل

  An TfrAt Missenses AlwAqE fy Aljz' AlxArjy mn brwtyn Almstqbl is the scope of the speculative cue يعتقد yEtqd and denoted by parentheses ().

Example 2:

وجدت الدراسة[(عدم]وجودفروق جوهرية)بين مجموعتي   الإصلاح باستخدام كلتا الجبيرتين.

Wjdt AldrAsp Edm [wjwd frwq jwhryp byn mjmwEty AlISlAH bAstxdAm kltA Aljbyrtyn

- عدمEdm is negation word(cue) and denoted by square brackets[].

- عدم وجود فروق جوهرية

  Edm wjwd frwq jwhryp is the scope of the negation cueعدمEdmand denoted by parentheses ().

VI. EXPERIMENTATION

In this corpus, five independent annotators annotated 1297 negative sentences and 1376speculative sentences, by ensuing the guidelinesthey provided by and using their linguistic experts.

The following tables (1, 2, and 3) show the main statistics about BioArabic corpus, number of document, negation sentences, speculation sentences and other information.

TABLE 1 STATISTICS ABOUT THE BIOARABIC CORPUS

Number Documents	70
Number of Sentences	10165
Number of Words	156236
Average length documents (in sentences)	125.54
Average of length documents (in words)	1929.54
Average of length sentences (in words)	15.37

TABLE 2 NEGATION STATISTICS IN THE BIOARABIC CORPUS

Number of Negation sentences	1297
The percentage of Negation sentences	0.127
Number of Negation cues	1567
Number of Words in scope	11856
Number of Scope	1342
Average of length scope	8.83

TABLE 3 SPECULATION STATISTICS IN THE BIOARABIC CORPUS

Number of Speculation sentences	1376
The percentage of Speculation sentences	0.135
Number of Speculation cues	1482
Number of Words in scope	12902
Number of Scope	1391
Average of length scope	9.27

VII. CONCLUSION AND FUTURE WORKS

To the best our knowledge this paper presents the first study in identification the negation and speculation cues and their scopes for Arabic biomedical texts. We reported on the creation of Arabic corpus for identification the negation and speculation cues and their scopes in biomedical texts. The corpus is freely available for research purposes for both computational linguistics and biologists, for Arabic texts. Compared to bioscope corpus, the size is appropriate and suitable for academic research needs. In future work, we are going to add new more practical features to our corpus and release the corpus to be accessible by any researcher. We will add new essential feature showed its effectiveness on detection the speculation and negation. This feature called by target [8],it is object described by a negative or speculative expression.

REFERENCES

[1] Cruz N., TaboadaM..andMitkovR., "A Machine Learning Approach to negation and Speculation Detection for Sentiment Analysis," the Journal of the Association for Information Science and Technology, 2015.

[2] VinczeV. and Hungary S., "Speculation and negation annotation in natural language texts: what thecase of BioScope might (not) reveal," Proceedings of the Workshop Negation and Speculation in Natural Language Processing (NeSp-NLP 2010), Uppsala, Sweden,2010.

[3] DesclésJ., MakkaouiO., HacèneT., "Automatic annotation of speculation in biomedical texts: new perspectives and large-scale evaluation," Proceedings of the Workshop Negation and Speculation in Natural Language Processing (NeSp-NLP 2010), Uppsala, Sweden,2010.

[4] Desclés J., Alrahabi M., and Desclés p."BioExcom: Automatic Annotation and categorization of speculative sentences in biological literature by a Contextual Exploration processing, " Proceedings of the 4th Language & Technology Conference (LTC), Poznań, Poland,2009.

[5] Kim J. and Park J.C.," Extracting Contrastive Information from Negation Patterns in Biomedical Literature," ACM Transactions on Asian Language Information Processing (TALIP), vol.5, no.1, pp.44-60,2006.

[6] Szarvas G., Vincze V., Farkas R., Csirik J.,"theBioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts," Proceedings of the WorkshopBioNLP ACL-2008, 2008.

[7] Kim J., OhtaT., and TsujiiJ.," Corpus annotation for mining biomedical events from literature,"BMC Bioinformatics, 2008.

[8] Zou B.,Zhou G.and Zhu Q., "Negation and Speculation Target Identification," Springer Berlin Heidelberg, pp. 34-45, 2014.

[9] Light M., Ying QiuX., and Srinivasan P.," The language of bioscience: Facts, speculations, and statements in between. In Lynette Hirschman and James Pustejovsky, Proceedings of the Workshop BioLINK 2004, Linking Biological Literature, Ontologies and Databases, pp. 17–24, Boston, Massachusetts, USA,2004.

[10] Dalianis H. and SkeppstedtM., "Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus", Proceedings of the Workshop Negation and Speculation in Natural Language Processing (NeSp-NLP 2010), Uppsala, Sweden, 2010.

[11] WiegandM., BalahurA., Roth B. and KlakowD., MontoyoA., "A Survey on the Role of Negation in Sentiment Analysis," Proceedings of the Workshop Negation and Speculation in Natural Language Processing (NeSp-NLP 2010), Uppsala, Sweden, 2010.

[12] AL-Khawaldeh F., “A Study of the Effect of Resolving Negation and Sentiment Analysis in Recognizing Text Entailment for Arabic," The World of Computer Science and Information Technology Journal (WSCIT), vol.5, Vol. 5, No. 7, 124-128, 2015