A UNIFIED FRAMEWORK TO IDENTIFY AND EXTRACT UNCERTAINTY CUES, HOLDERS, AND SCOPES IN ONE FELL-SWOOP

BY

RANIA MOSTAFA AL-SABBAGH

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Associate Professor Roxana Girju, Chair and Co-Director of Research
Assistant Professor Jana Diesner, Co-Director of Research
Professor Elabbas Benmamoun
Associate Professor Julia Hockenmaeir

ABSTRACT

Uncertainty refers to the language aspects that express hypotheses and speculations where propositions are held as (un)certain, (im)probable, or (im)possible. Automatic uncertainty analysis is crucial for several Natural Language Processing (NLP) applications that need to distinguish between factual (i.e. certain) and nonfactual (i.e. negated or uncertain) information. Typically, a comprehensive automatic uncertainty analyzer has three machine learning models for uncertainty detection, attribution, and scope extraction. To-date, and to the best of my knowledge, current research on uncertainty automatic analysis has only focused on uncertainty attribution and scope extraction, and has typically tackled each task with a different machine learning approach. Furthermore, current research on uncertainty automatic analysis has been restricted to specific languages, particularly English, and to specific linguistic genres, including biomedical and newswire texts, Wikipedia articles, and product reviews.

In this research project, I attempt to address the aforementioned limitations of current research on automatic uncertainty analysis. First, I develop a machine learning model for uncertainty attribution, the task typically neglected in automatic uncertainty analysis. Second, I propose a unified framework to identify and extract uncertainty cues, holders, and scopes in one-fell swoop by casting each task as a supervised token sequence labeling

problem. Third, I choose to work on the Arabic language, in contrast to English, the most commonly studied language in the literature of automatic uncertainty analysis. Finally, I work on the understudied linguistic genre of tweets.

This research project results in a novel NLP tool, i.e., a comprehensive automatic uncertainty analyzer for Arabic tweets, with a practical impact on NLP applications that rely on uncertainty automatic analysis. The tool yields an F_1 score of 0.759, averaged across its three machine learning models. Furthermore, through this research, the research community and I gain insights into (1) the challenges presented by Arabic as an agglutinative morphologically-rich language with a flexible word order, in contrast to English; (2) the challenges of the linguistic genre of tweets for uncertainty automatic analysis; and (3) the type of challenges that my proposed unified framework successfully addresses and boosts performance for.

To Azza, Mostafa, and Muhammad.

ACKNOWLEDGMENTS

I would have never been able to finish my thesis without the love and support of my mother, Azza, my father, Mostafa, and my brother, Muhammad. They have always been there cheering me up and standing by me through the good and the not-so-good times.

I would also like to thank Roxana Girju and Jana Diesner for their continuous feedback and support. Finally, I really appreciate the feedback I received from my committee members: Elabbas Benmamoun and Julia Hockenmaier.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Uncertainty refers to the language aspects that express hypotheses and speculations where propositions are held as (un)certain, (im)probable, or (im)possible. Different terms have been used to refer - more or less - to the same concept: committed belief (Diab et al., 2009), epistemic modality (Palmer, 1986), evidentiality (Aikhenvald, 2004), factuality (Saurí and Pustejovsky, 2009), speculation (Apostolova et al., 2011; Vincze et al., 2011; Vlachos and Craven, 2010), and veridicality (de Marneffe et al., 2012). Automatic uncertainty analysis is crucial for several Natural Language Processing (NLP) applications that need to distinguish between factual (i.e. certain) and nonfactual (i.e. negated or uncertain) information, including:

- Rumor detectors that identify statements with unverified truth values (Qazvinian et al., 2011).
- Credibility analyzers that detect disinformers who endorse rumors and further spread them (Castillo et al., 2011; Soni et al., 2014).
- Question-answering systems that evaluate the truth value of Web-based information to use as answers (Azari et al., 2003; de Marneffe et al., 2009).

- Medical text analyzers that decide whether a given patient definitely suffers, probably suffers, or does not suffer from an illness (Szolovits, 1995; Mowery et al., 2012).

A comprehensive automatic uncertainty analyzer typically comprises three machine learning models for:

- **Uncertainty detection**, i.e., the identification and extraction of uncertainty linguistic cues.
- **Uncertainty attribution**, i.e., the identification and extraction of uncertainty holders.
- **Uncertainty scope extraction**, i.e., the identification and extraction of uncertainty propositions.

Current research on automatic uncertainty analysis has been limited to uncertainty detection and scope extraction, whereas attribution has been either ignored (Baker et al., 2012); or simplistically handled by either setting the text writer as the default holder (Diab et al., 2009) or using a predefined set of prototypical holders (Wiegand and Klakow, 2011a). One reason for the less interest in uncertainty attribution is related to the types of linguistic genres typically studied for automatic uncertainty analysis. Most current research on uncertainty automatic analysis has focused on biomedical and newswire texts (Li et al., 2010; Szarvas and Gurevych, 2013; Szarvas et al., 2008; Vlachos and Craven, 2010), Wikipedia articles (Tjong and Sang, 2010; Vincze, 2013), and product reviews (Díaz, 2013) in which uncertainty can be ascribed to the text writer in most cases unless there is a quote. Even for quotes, in many cases holders are limited to such prototypical noun phrases as *experts*, *analyzers*, and *scientists* (Wiegand and Klakow, 2011a). As a result, no much attention has been given to uncertainty attribution. Besides, some researchers have just decided to dedicate more time and effort to uncertainty detection and scope extraction for no clear reasons (Baker et al., 2012; Wei et al., 2013). With a highly interactive linguistic genre such as tweets, default and prototypical holders are unlikely to work: Twitter users not only post about their own uncertainties, but also they post about the uncertainties they think others may have, whether those others are their own followers or some other third parties. As a result, uncertainty attribution is indispensable for automatic uncertainty analysis in tweets, especially if I want to use uncertainty automatic analysis to detect rumor spreaders or to rank Twitter users based on their credibility. The diversity of uncertainty holders makes the linguistic genre of tweets specially interesting. Another reason that makes tweets interesting is that nowadays Twitter is one of the fastest growing information sources. As a result, it is crucial to develop machine learning models that can process Twitter-based information both accurately and fast.

Current research on automatic uncertainty analysis has also been limited in terms of its approaches to uncertainty detection and scope extraction. None of the research that tackles the two uncertainty-related tasks simultaneously uses a unified framework. For example, Ørelid et al. (2010), Yang et al. (2012), Apostolova et al. (2011), and Velldal et al. (2010) use a token sequence labeling approach to uncertainty detection and then a rule-based approach to uncertainty scope extraction. Zhao et al. (2010) define both tasks as token sequence labeling problems; yet use separate feature sets for each task and do not use the output of one task to inform the other. The absence of a unified framework for all uncertainty-related tasks consumes more time for feature extraction and deprives tasks from informing one another.

Current research on automatic uncertainty analysis has also been limited to specific languages. There is a plethora of work on English (Matsuyoshi et al., 2010; Prabhakaran, 2010; Rubin, 2007; Rubinstein et al., 2013; Ruppenhofer and Rehbein, 2012; Saurí and Pustejovsky, 2009; Szarvas et al., 2008; Tang et al., 2010; Vincze et al., 2011; Wei et al., 2013), French (Goujon, 2009), Portuguese (Hendrickx et al., 2012; Avila and Mello, 2013), and Swedish (Mowery et al., 2012), yet nothing for agglutinative morphologically-rich languages except for Hungarian (Vincze, 2014). One reason for the little research on agglutinative morphologically-rich languages with a flexible word order is the lack of uncertainty-annotated corpora for such languages. Another reason is the significant challenges that those languages present for standard approaches designed for English. First, agglutination and productive inflectional morphology can package much uncertainty information in the same token; so that information about the holder can be encoded in the morphological inflection of the cue; and a single token can be the uncertainty cue and encode information about the scope simultaneously. Second, morphological richness introduces a high number of variable tokens, leading to data sparsity. Finally, flexible word order results in very long dependencies that are not usually present in the English language. As a result, only Vincze (2014) has worked on uncertainty automatic analysis for agglutinative morphologically-rich languages with a flexible word order, represented by Hungarian; yet, she has limited her work to uncertainty detection.

In this research project, therefore, I attempt to address the aforementioned limitations as I build a comprehensive automatic uncertainty analysis system for Arabic tweets with three machine learning models for uncertainty detection, attribution, and scope extraction. The novelty of the research does not only lie in working on an understudied uncertainty task, i.e., attribution, an understudied linguistic genre, i.e., tweets, and an understudied language, i.e., Arabic, but also in my proposed unified framework to build the three aforementioned machine learning models. I propose a unified framework to identify and extract uncertainty cues, holders, and scopes in one fell-swoop, by casting each task as a supervised sequence labeling machine learning problem. The three uncertainty-related tasks are organized in a pipeline fashion starting with uncertainty detection, followed uncertainty attribution, and then scope extraction, for each identified cue, one cue at a time. The unified framework relies on Support Vector Machines and a large set of morphological, syntactic, lexical, semantic, pragmatic, dialectal, and genre-specific features. Many features are shared across the three machine learning models; and hence, the time needed for feature extraction is reduced to speed up the system. Furthermore, the predictions of one model inform the predictions of the next model in the pipeline. For instance, once a token is predicted as encoding an uncertainty holder, it is excluded from candidate tokens for uncertainty scopes because a single token cannot encode information about uncertainty holders and scopes at the same time. This exclusion process boosts performance, especially that tweets are typically short texts with a few tokens.

The rest of this thesis comprises six chapters. Chapter 2 describes my uncertainty-annotated corpus, its annotation guidelines, results, and disagreement factors. Chapters 3, 4, and 5, describe my three machine learning models for uncertainty detection, attribution, and scope extraction, respectively, with information about classification features, experimental setup and results, errors, and comparisons with others' related work. Finally, Chapter 6 wraps up this entire thesis and highlights the most important findings and future work directions.

CHAPTER 2

CORPUS DEVELOPMENT

2.1 Introduction

As I mentioned in Chapter 1, one reason for the little research on automatic uncertainty analysis in agglutinative morphologically-rich languages with a flexible word order is the lack of uncertainty-annotated corpora to train, test, and evaluate supervised machine learning models. As a result, my first step for this research project is to develop such a corpus for Arabic tweets. The corpus is a novel NLP resource that I share with the community to trigger further research into the understudied areas of uncertainty attribution, uncertainty automatic analysis for the linguistic genre of tweets and for agglutinative morphologically-rich languages with a flexible word order [1].

The remainder of this chapter is organized as follows: Section 2.2 describes raw corpus harvesting, annotation scheme and results, the inter-annotator disagreement factors, and statistics for the final annotated corpus; and Section 2.3 compares and contrasts my corpus with others' related work.

[1] All resources developed throughout this research project are available at: www.rania-alsabbagh.com

2.2 Corpus Development

2.2.1 Raw Corpus Harvesting and Description

I harvested 21,716 tweets with REST Twitter API from June 2012 to June 2013. As search queries, I used such hashtags as: #Egypt [2], #Jan25, #Ikhwan (i.e. the Muslim Brotherhood), #مرسي *mrsy* (Morsi)[3] , #الدستور *#Aldstwr* (the constitution), and #السيسي *#Alsysy* (Elsisi), among many others. Most of the used hashtags are of political interest. This sets the domain of the harvested corpus to politics.

The harvested corpus comes from a variety of Twitter users: (1) press users such as newspapers, TV stations, and political campaigns; and (2) individual users, including politicians, journalists, political activists, and other ordinary people. This entails that the corpus contains more than one variety of the Arabic language. The first variety is Modern Standard Arabic (MSA), which is the formal variety of Arabic mostly used by the press users. The other varieties include several local Arabic dialects used mostly by the individual users. A quick look at the harvested corpus shows that the predominant Arabic dialect is Egyptian Arabic (EA). This is expected given that most of the hashtags used for corpus harvesting discuss political issues of interest in Egypt.

The final harvested corpus contains 521,786 word tokens and 64,445 word types, where words are defined as strings delimited by white spaces.

[2] English hashtags can be used with Arabic tweets.

[3] Buckwalter's Arabic Transliteration Scheme: http://www.qamus.org/transliteration.htm

2.2.2 Annotation Tasks

Our annotation scheme comprises three tasks to annotate tokens encoding uncertainty cues, holders, and scopes.

2.2.2.1 Annotating Uncertainty Cues

For the first annotation task, annotators are required to identify and extract Uncertainty Cues (UCs) in each given tweet. Annotators are informed that:

- UCs express hypotheses and speculations where propositions are held as (un)certain, (im)probable, or (im)possible.
- UCs are synonymous to our given set of prototypical UCs.
- UCs can be nominals, verbs, or particles.
- UCs can be unigrams or multiword expressions.
- The components of multiword UCs are not necessarily adjacent.

The first guideline depicts my generic definition of uncertainty that is not geared to any specific NLP application or task. This guideline is further supported by my second guideline which casts uncertainty annotation as a synonymy judgement task. This enables annotators to check whether each tokens they mark as a UC truly confines to my definition in the first guideline or not. This simplification of UC annotation as a synonymy judgement task increases inter-annotator reliability rates and gives the annotators the chance to revise their initial annotations as I show in my preliminary annotation study (Al-Sabbagh et al., 2014a). The set of prototypical UCs comprises unambiguous affirmative and negative UCs of different Arabic dialects, parts of speech, and morphological inflections. Table 2.1 shows my set [4].

The third guideline for this task instructs annotators that Arabic UCs, like in many other languages including English, come in different parts of speech: nominals, verbs, and particles. Examples of nominal UCs are the common noun احتمالات *AHtmAlAt* (gloss: possibilities; English: there is a possibility), the present participle شايف *$Ayf* (gloss: thinks.3.sg.msc.imprf; English: he thinks), the adjective مؤكد *m&kd* (gloss: sure.3.sg.msc; English: surely), and the adverb ربما *rbmA* (gloss: probably; English: probably), in examples 1-4[5] [6], respectively. In example 5, لازم *lAzm* (gloss: must; English: must) is an example of auxiliary verb UCs. Lexical verb UCs are like يعتقد *yEtqd* (gloss: thinks.3.sg.msc.imprf; English: he thinks) in example 6. The emphatic particle قد *qd* (gloss: may; English: may) in example 7 illustrates particle UCs.

1. - هناك احتمالات مواجهة بين البحرين وإيران.
   - ***hnAk AHtmAlAt*** *mwAjhp byn AlbHryn w <yrAn*.
   - **there possibilities** confrontation.sg.fm between Bahrain and-Iran.
   - **There is a possibility for** a confrontation between Bahrain and Iran.

[4] For every example, (1) I give the Arabic version, (2) Buckwalter's transliteration, (3) a gloss that mirrors Arabic morphology, semantics, and syntax, and (4) an English translation. For example, **think.1.sg.imprf** means the Arabic equivalent of the verb *think* inflected for the first person singular in the imperfective aspect; **think.3.sg.msc.prf** means the Arabic equivalent of the verb *think* inflected for the third masculine singular person in the perfective aspect; **the-most** means the definite article attached as a proclitic to the superlative adjective *most*.

[5] For each example, the 1st line is the original Arabic tweet. The 2nd line is the Buckwalter's transliteration. The 3rd line is a gloss. The 4th line is an English translation.

[6] For all examples, uncertainty cues are in boldface; uncertainty holders are double underlined; and uncertainty scopes are single underlined.

Arabic	Trans.	Gloss	English	POS	Polarity
طبعا	*TbEA*	definitely	definitely	adverb	affirmative
حتما	*HtmA*	absolutely	absolutely	adverb	affirmative
ربما	*rbmA*	probably	probably	adverb	affirmative
على الأغلب	*ElY Al>glb*	on the-most	most probably	mwe	affirmative
أعتقد	*>Etqd*	think.1.sg.imprf	I think	verb	affirmative
نعتقد	*nEtqd*	think.1.pl.imprf	we think	verb	affirmative
يعتقد	*yEtqd*	thinks.3.sg.msc.imprf	he thinks	verb	affirmative
يعتقدون	*yEtqdwn*	think.3.pl.imprf	they think	verb	affirmative
معتقدش	*mEtqd$*	not-think.1.sg.imprf-not	I do not think	verb	negative
منعتقدش	*mnEtqd$*	not-think.1.pl.imprf-not	we do not think	verb	negative
ميعتقدوش	*myEtqdw$*	not-think.1.pl.imprf-not	they do not think	verb	negative
متهيألي	*mthy>ly*	think.1.sg.imprf	I guess	verb	affirmative
متهيألنا	*mthy>lnA*	think.1.pl.imprf	we guess	verb	affirmative
متهيألكم	*mthy>lkm*	think.2.pl.imprf	you guess	verb	affirmative
متهيأله	*mthy>lh*	thinks.3.sg.msc.imprf	he guesses	verb	affirmative
متهيألها	*mthy>lhA*	thinks.3.sg.fm.imprf	she guesses	verb	affirmative
متهيألهم	*mthy>lhm*	think.3.pl.imprf	they guess	verb	affirmative
ظننت	*Znnt*	thought.1.sg.prf	I thought	verb	affirmative
ظننا	*ZnnA*	thought.1.pl.prf	we thought	verb	affirmative
ظنت	*Znt*	thought.3.sg.fm.prf	she thought	verb	affirmative
لم يعرف	*lm yErf*	not knows.3.sg.msc.imprf	he did not know	verb	negative
لم يدرك	*lm ydrk*	not realizes.3.sg.msc.imprf	he did not realize	verb	negative
لم يتصور	*lm ytSwr*	not imagines.3.sg.msc.imprf	he did not imagine	verb	negative
لا يعرف	*lA yErf*	not knows.3.sg.msc.imprf	he does not know	verb	negative

لا يدرك	*lA ydrk*	not realizes.3.sg.msc.imprf	he does not realize	verb	negative
لا يتصور	*lA ytSwr*	not imagines.3.sg.msc.imprf	he does not imagine	verb	negative
عارفين	*EArfyn*	knowing.pl.imprf	we/you/they know	present participle	affirmative
متأكد	*mt>kd*	sure.sg.imprf	I/he am/is sure	adjective	affirmative
واثقين	*wAvqyn*	confident.pl.imprf	we/you/they are confident	present participle	affirmative
شاكين	*$Akyn*	doubting.pl.imprf	we/you/they doubt	present participle	negative
احتمال	*AHtmAl*	possibility	there is a possibility	noun	affirmative
على يقين	*ElY yqyn*	on confidence	confident	mwe	affirmative
من المستبعد	*mn AlmstbEd*	from the-excluded	it is unlikely	mwe	negative
من غير الوارد	*mn gyr AlwArd*	from not the-likely	it is unlikely	mwe	negative
من المتوقع	*mn AlmtwqE*	from the-expected	it is expected	mwe	affirmative
من المستحيل	*mn AlmstHyl*	from the-impossible	it is impossible	mwe	negative
من غير المحتمل	*mn gyr AlmHtml*	from not the-possible	it is not possible	mwe	negative
من غير المنتظر	*mn gyr AlmntZr*	from not the-waited	it is unexpected	mwe	negative
قد	*qd*	may/indeed	may/indeed	particle	affirmative

Table 2.1: Prototypical UCs

2.
- محامي #مرسي كان شايف إن ٨٥ % من مية النيل بيتيجي من بحيرة فيكتوريا.
- *mHAmy #mrsy* ***kAn $Ayf <n*** *85% mn myp Alnyl btyjy mn bHyrp fyktwryA*.
- laywer #Morsi **was.3.sg.msc.prf thinking.3.sg.msc.prf that** 85% of water the-Nile coming.3.sg.fm.imprf from Lake Victoria.
- #Morsi's lawyer **thought that** 85% of the Nile comes from Lake Victoria.

3.
- مؤكد مرسي حينجح.
- ***m&kd*** *mrsy HynjH*.
- **sure.sg.msc** Morsi will-win.3.sg.msc.imprf.
- **Surely,** Morsi will win.

4.
- ربما الجيش كان مستني الفرصة للاستيلاء على الحكم.
- ***rbmA*** *Aljy$ kAn mstny AlfrSp llAstylA' ElY AlHkm*.
- **Maybe**, the-army was.3.sg.msc.prf waiting.3.sg.msc.imprf the-chance to-the-seize on the-power.
- **Maybe**, the army was waiting for a chance to seize power.

5.
- العساكر دول لازم رايحين على العباسية.
- *AlEsAkr dwl* ***lAzm*** *rAyHyn ElY AlEbAsyp*.
- the-police those **must** heading.3.pl.imprf to Abbasia.
- Those riot police forces **must** be heading to Abbasia.

6.
- البعض يعتقد أن هناك مؤامرة من المجلس العسكري.
- *AlbED* ***yEtqd >n*** *hnAk m&Amrp mn Almjls AlEskry*.
- some **think.3.sg.msc.imprf that** there conspiracy from the-council the-military.
- Some people **think that** the Military Council is conspiring.

7.
- قد تأخذ كتابة الدستور أكثر من شهرين.

- ***qd*** *t>xz ktAbp Aldstwr >kvr mn $hryn*.
- **may** takes.3.sg.fm.imprf writing the-constitution more than months.dual.
- The constitution write-up **may** take more than two months.

Like in many other languages, Arabic UCs can be either unigrams or multiword expressions. In my preliminary study (Al-Sabbagh et al., 2014b), I have identified three types of Arabic multiword UCs based on their morpho-syntactic and lexical flexibility:

- **Type 1** comprises idiomatic expressions such as لعل وعسى *lEl wEsY* (gloss: may and-hopefully; English: maybe) and فيما يبدو *fymA ybdw* (gloss: in-what seems.3.sg. msc.imprf; English: seemingly). They are morphologically, syntactically, and lexically fixed. Furthermore, they do not allow insertions in-between their boundaries.

- **Type 2** includes morphologically and syntactically productive multiword UCs such as: يتأكد من أن *yt>kd mn >n* (gloss: confirms.3.sg.msc.imprf from that; English: he makes sure that). They inflect for gender, person, number, and aspect; hence, they are morphologically productive. Furthermore, they are syntactically productive because they allow several linguistic constituents to be inserted in-between their boundaries. These linguistic constituents can be:

  - Adverbial phrases such as تماما *tmAmA* (gloss: completely; English: completely) in as in واثق تماما في *wAvq tmAmA fy* (gloss: confident.3.sg.msc completely in; English: is completely confident that).

  - Noun phrases such as الثوار *AlvwAr* (gloss: the-revolutionists; English: the revolutionists) in يتأكد الثوار من أن *yt>kd AlvwAr mn >n* (gloss: confirms.3.sg.

msc.imprf the-revolutionists from that; English: the revolutionists make sure that).

- Prepositional phrases such as بشدة *b$dp* (gloss: with-intensity; English: very much) in أشك بشدة أن *>$k b$dp >n* (gloss: doubt.1.sg.imprf; English: I very much doubt that).

- **Type 3** comprises lexically and syntactically productive multiword UCs such as -من المتوقع أن *mn AlmtwqE >n* (gloss: from the-expected that; English: it is expected that). They are lexically productive in the sense that their lexical meaning and uncertainty degree rely on their head word. Thus, with the same aforementioned syntactic structure, we can have من المؤكد أن *mn Alm&kd >n* (gloss: from the-confirmed that; English: it is confirmed that), من المرجح أن *mn AlmrjH >n* (gloss: from the-probable that; English: it is probable that), and من المحتمل أن *mn AlmHtml >n* (gloss: from the-possible that; English: it is possible that), among many others. Meanwhile, they are syntactically productive because they allow adverbial, noun, and prepositional phrases within their boundaries as in Type 2.

As the aforementioned types of multiword UCs suggest, Arabic recognizes both continuous and discontinuous multiword UCs. For discontinuous UCs, annotators are instructed that they have to mark the entire UC including all intervening linguistic constituents. Thus, in example 8, the UC is من غير الوارد أن *mn gyr AlwArd >n* (gloss: from not the-likely that; English: it is unlikely that), with the negation particle included.

8. #غزلان: من غير الوارد أن يدعم الإخوان #أبو الفتوح في سباق الرئاسة. •

- *#gzlAn*: ***mn gyr AlwArd >n*** *ydEm Al<xwAn #>bw AlftwH fy sbAq Alr}Asp*.
- #Ghizlaan: **from not the-likely that** supports.3.sg.msc.imprf Ikhwan #Abu Alfotoh in competition the-presidency.
- #Ghizlaan: **it is unlikely that** the Muslim Brotherhood will vote for Abulfotoh for presidency.

2.2.2.2 Annotating Uncertainty Holders

For my second annotation task, annotators are required to identify Uncertainty Holders (UHs) and extract the linguistic constituents, typically noun phrases/clauses, that correspond to them, if applicable. In my corpus, there are three possible types of UHs:

- **Type 1** comprises UCs whose holders are the same as the users who posted the uncertainty-laden tweets. Type 1 is encoded by:

  - 1st person pronouns such as انا *AnA* (gloss: I; English: I) in example 9:

    9. - انا متهيالي محدش هيشارك.
       - *AnA* ***mthyAly*** *mHd$ hy$Ark*.
       - I **think.1.sg.imprf** not-one-not will-participate.3.sg.msc.imprf.
       - I **think** no one will participate.

  - UCs morphologically inflected for the 1st person as in مظنش *mZn$* (gloss: not-think.1.sg.imprf-not; English: I do not think) in example 10:

    10. - مظنش اصوت لمرسي.
        - ***mZn$*** *ASwt lmrsy*.

        - **not-think.1.sg.imprf-not** vote.1.sg.imprf for-Morsi.
        - I **do not think** I will vote for Morsi.

    - impersonal and passive voice UCs such as من المنتظر أن *mn AlmntZr >n* (gloss: from the-expected that; English: it is expected that) in example 11:

        11.
            - من المنتظر أن يلقي الرئيس خطابا اليوم.
            - ***mn AlmntZr >n*** *ylqy Alr}ys xTAbA Alywm.*
            - **from the-waited that** gives.3.sg.msc.imprf the-president speech.sg the-day.
            - **It is expected that** the president will give a speech today.

- **Type 2** comprises UCs whose holders are the followers of the users who posted the uncertainty-laden tweets. Type 2 is encoded by:

    - 2nd person pronouns such as أنت *>nt* (gloss: you.sg.msc; English: you) in example 12:

        12.
            - انت شايف مرسي أسد وأنا شايفه حمار.
            - *Ant* ***$Ayf*** *mrsy Asd wAnA.* ***$Ayf*** *h HmAr.*
            - you.sg.msc **think.2.sg.msc.imprf** Morsi lion and-I **think.1.sg.msc.imprf**-him donkey.
            - You **think** Morsi is as brave as a lion, but I **think** he is as stupid as a donkey.

    - UCs morphologically inflected for the 2nd person like افتكرتوا *AftkrtwA* (gloss: thought.2.pl.prf; English: you thought) in example 13:

        13.
            - افتكرتوا الإخوان هيسكتوا بعد اقتحام مقراتهم.

- ***AftkrtwA*** *Al<xwAn hysktwA bEd AqtHAm mqrAthm.*
- **thought.2.pl.prf** the-brotherhood will-remain.silent.3.pl.imprf after breaking.into headquarters-their.
- You **thought that** the Muslim Brotherhood members will not react against breaking into their headquarters.

- **Type 3** comprises UCs whose holders are neither Type 1 or Type 2 and are expressed as:
  - noun phrases/clauses of directly cited holders such as صباحي *SbAHy* (gloss Sabbahy; English: Sabbahy - a masculine proper noun) in example 14:

    14.
    - صباحي: لدي شكوك بأن حادث السيارة كان مدبرا.
    - *sbAHy*: ***ldy $kwk b>n*** *HAdv AlsyArp kAn mdbrA.*
    - Sabbahy: **for-me doubts with-that** accident.sg.msc the-car.sg.fm was.3.sg.msc.prf premeditated.sg.msc.
    - Sabbahy: I **suspect that** the car accident was premeditated.

  - noun phrases/clauses of indirectly cited holders such as أحمد مكي *>Hmd mky* (gloss: Ahmed Mekky; English: Ahmed Mekky - masculine proper nouns) in example 15:

    15.
    - أحمد مكي بيقول أنه ما يعرفش أن فيه ١١٤ طفل محبوسين.
    - *>Hmd mky byqwl >nh* ***mAyErf$ >n*** *fyh 114 Tfl mHbwsyn.*
    - Ahmed Mekky saying.3.sg.msc.imprf that-he **not knows.3.sg.msc.imprf-not that** there 114 kid arrested.pl.

- Ahmed Mekky says that he **does not know that** there are 114 arrested kids.

– **3rd** person pronouns such as هو *hw* (gloss: he; English: he) in example 16:

16.
- هو **مش مدرك ان** في اماكن مينفعش يتقطع عنها النور.
- *hw **m$ mdrk An** fyh AmAkn mynfE$ ytqTE EnhA Alnwr.*
- he **not realizing.3.sg.msc.imprf that** there places not-can-not cut.3.sg.msc .imprf.passive off-them the-electricity.
- He **does not realize that** there are places that cannot have an electricity outage.

The aforementioned description of the types of UHs in my corpus shows that the productive inflectional morphology of Arabic can package more than one piece of information pertinent to uncertainty into a single token. As a result, مظنش *mZn$* (gloss: not-think.1.sg.imprf-not; English: I do not think) in example 10 and افتكرتوا *AftkrtwA* (gloss: thought.2.pl.prf; English: you thought) in example 13 are not only UCs but also they comprise UH information. That type of challenge has not been considered in earlier work on uncertainty annotation and automatic analysis because most earlier work focuses on the English language in which cues and holders are typically represented by separate tokens.

Annotators are instructed that if the UH information is encoded in the morphology or the semantics of the UC as in example 11, the text segment of the UC is to be the same for the UH. Otherwise, if UHs are represented by separate tokens, annotators are instructed to follow the maximal length principle from (Szarvas et al., 2008) so that the marked text segment comprises all complements and adjuncts pertinent to the UH. According to the maximal length principle, the UH in example 17 is صباحي للمصري اليوم *SbAHy llmSry*

Alywm (gloss: Sabbahy for-Almasry Alyoum; English: Sabbagh for Almasry Alyoum[7]) not only صباحي *SbAhy* (gloss: Sabbahy; English: Sabbahy).

17.
- صباحي للمصري اليوم: يمكن مرسي حابب يرمضن مع الجنزوري.
- *SbAHy llmSry Alywm*: ***ymkn*** *mrsy HAbb yrmDn mE Aljnzwry*.
- Sabbahy for-Almasry Alyoum: **maybe** Morsi wants.3.sg.msc.imprf spend.Ramadan .3.sg.msc.imprf with Elganzoury.
- Sabbahy for Almasry Alyoum: **maybe** Morsi wants to spend Ramadan with Elganzoury.

2.2.2.3 Annotating Uncertainty Scopes

Uncertainty scope annotation identifies the text segments that encode the propositions modified by the UCs. Annotators are instructed to use the same maximal length principle used for annotating UHs so that the marked text segment includes all the complements and adjuncts related to the scope.

Scope annotation involves several challenges that annotators are instructed to deal with. Nesting, where a UC and its scope are embedded in another UC's scope, is common. In example 18, مقتنعين *mqtnEyn* (gloss: convinced.pl; English: are convinced) and its scope بجدوى الحوار *bjdwY AlHwAr* (gloss: with-usefulness the-talk; English: that the talk is useful) are both embedded in the scope of مظنش *mZn$* (gloss: not-think.1.sg.imprf-not; English: I do not think).

18.
- مظنش انهم مقتنعين بجدوى الحوار.
- ***mZn$ An**hm **mqtnEyn b**jdwY AlHwAr*.

[7] An Egyptian newspaper

- **not-think.1.sg.imprf-not that-**they **convinced.pl with-**usefulness the-talk.
- I **do not think** they **are convinced that** it is important to talk.

A single UC may have one or more scopes. In example 19, the scope of بيتهيألهم *bythy>lhm* (gloss: imagine.3.pl.imprf; English: they imagine) comprises two coordinating complement clauses.

19.

- أولادنا بيتهيألهم إن دم أخواتهم راح هدر وإن عندهم ثأر مع السلطة.
- *>wlAdnA* ***bythy>lhm*** *<**n** dm >xwAthm rAH hdr w<**n** Endhm v>r mE ElslTp*.
- kids-our **imagine.3.pl.imprf that** brothers-their blood went.3.sg.msc.prf in.vain and-**that** have.3.pl.imprf-they revenge with the-authorities.
- Our kids **imagine that** their brothers were killed in vain and **that** they have to take revenge from the authorities.

Two or more UCs - typically conjoined by a coordinating conjunction - can share the same scope. In example 20, the coordinating UCs فاهم وعارف *fAhm wEArf* (gloss: knows.3.sg.msc.imprf and-understands.3.sg.msc.imprf; English: knows and understands) share the same complement clause scope.

20.

- الشعب فاهم وعارف انه من اذناب النظام #شفيق.
- _Al$Eb_ **_fAhm wEArf An_**_h mn A*nAb AlnZAm. #$fyq_
- the-people **knows.3.sg.msc.imprf** and **understands.3.sg.msc.imprf that-**he from followers the-regime. #Shafiq
- The people **know and understand that** he is part of the old regime. #Shafiq

Scopes are not necessarily adjacent to their UCs. In example 21, the scope starts three words to the right of the UC بقتنع *bAqtnE* (gloss: convinced.1.sg.imprf; English: get convinced) given that the adverbial phrase اكتر واكتر *Aktr wAktr* (gloss: more and-more; English: more and more) falls in-between the UC and its scope.

21.

- كل يوم باقتنع اكتر واكتر اننا كنا محتاجين ديكتاتور وطني عادل.
- *kl ywm* ***bAqtnE*** *Aktr wAktr* ***An****nA knA mHtAjyn dyktAtwr wTny EAdl.*
- every day **convinced.1.sg.imprf** more and-more **that**-we were.1.pl.prf needing.pl dictator.sg.msc patriotic.sg.msc fair.sg.msc.
- Every day, I **get** more and more **convinced that** we needed a patriotic and fair dictator.

Scopes can precede, follow, or surround their UCs. Many of the aforementioned examples have the scopes following their UCs. In example 22, the scope surrounds its UC (i.e. يبدو *ybdw* (gloss: seems.1.sg.msc.imprf; English: seems)). In example 23, the scope precedes its UC فيما يبدو *fymA ybdw* (gloss: in-what seems.3.sg.msc.imprf; English: seemingly). Annotators are instructed to mark all scope-encoding tokens, including complements and adjuncts, following the maximal length principle; whether those scopes are continuous or discontinuous (i.e. they are interrupted by their UCs).

22.

- وعود مرسي ليست فيما يبدو دينا عليه.
- *wEwd mrsy lyst* ***fymA ybdw*** *dynA Elyh.*
- Morsi promises not **in-what seems.3.sg.msc.imprf** debt on-him.
- Morsi's promises are not doable, **seemingly**.

23.

- حملة تشويه ثورة يناير وإعادة عقارب الساعة للوراء بدأت فيما يبدو.
- *Hmlp t$wyh vwrp ynAyr bd>t* ***fymA ybdw*** .
- campaign distortion Revolution January started.3.sg.fm.prf **in-what seems.3 .sg.msc.imprf**.

- A campaign to distort the image of January’s revolution has started, **seemingly**.

2.2.3 Annotation Results

I hired two annotators, both of whom are native speakers of Arabic and have studied linguistics at a university level. I had the annotators extensively trained on my guidelines, before asking each one of them to work independently on the entire corpus.

To measure inter-annotator reliability, I use α Coefficient (Krippendorff, 2004) that relies on the difference between observed and expected disagreement, unlike Kappa κ Coefficient that works on observed and expected agreement. The main advantages of α Coefficient are that:

- It is suitable for any number of annotators.
- It can be used for any size of data and with missing or incomplete data.
- It is easily interpretable as 1 means perfect reliability and 0 means no reliability.

For those reasons, α Coefficient is used to measure inter-annotator reliability in the most recent work on uncertainty annotation, including Rubinstein et al. (2013), Cui and Chi (2013), and my own preliminary work to create this Arabic uncertainty-annotated corpus (Al-Sabbagh et al., 2014a). For a full comparison between α, κ, and other coefficients, we refer the reader to Artstein and Poesio (2008). Table 2.2 shows the inter-annotator reliability rates per annotation task.

I finally acted as an adjudicator to settle the disagreements between the two annotators and come up with the final corpus. The final annotated corpus comprises 7,461 out of the

Task	Alpha
Cues	0.930
Holders	0.856
Scopes	0.825

Table 2.2: Inter-annotator reliability alpha rates per annotation task

initially-harvested 21,716 tweets that do not have any uncertainty information. The rest of the tweets comprise 17,317 uncertainty cues, of which 3,697 are unigrams and 13,620 are multiword expressions. Each cue is annotated for holders, of which 7,992 are encoded in the morphological inflections of the cues, whereas the rest are represented via personal pronouns or any other base or complex noun phrases. Each cue is also annotated for scopes, of which 3,857 are discontinuous[8].

2.2.4 Discussion and Disagreement Analysis

One reason for my high inter-annotation reliability rates is my set of prototypical UCs. It helps annotators avoid annotation inconsistencies resulting from their own personal interpretations of what hypotheses and speculations are. Another reason is the simplicity of the genre of tweets, where sentences are typically short and straightforward. Short sentences entail that annotators do not have many linguistic constituents to annotate and not many choices per annotation task; besides, long dependencies between UCs and their scopes become less common with short sentences. Furthermore, the majority of the uncertainty in my corpus is held by the Twitter users themselves, which suggests

[8] The corpus is available at www.rania-alsabbagh.com

that Twitter users are more likely to post about their own uncertainties than about others'. Type 1 UHs are the easiest to label. Hence, the inter-annotator reliability rate for annotating UHs is considerably high.

For each annotation task, I manage to identify the major disagreement factors. For UCs, disagreement is mainly attributed to opinionated evidential and highly polysemous UCs. Opinionated evidential UCs, such as يزعم *yzEm* (gloss: claims.3.sg.msc.imprf; English: he claims) and يدعي *ydEy* (gloss: alleges.3.sg.msc.imprf; English: he alleges), do not only mark reported speech, but also they communicate the reporter's own speculation about the truth value of the reported proposition. They entail that from the reporter's perspective the proposition is nonfactual (i.e. false or uncertain). Hence, annotators disagree as to whether these cues should be labeled as UCs or not.

Highly polysemous UCs, such as يمكن *ymkn* (gloss: enables.3.sg.msc.imprf/can/possible; English: he enables/can/it is possible) and يقسم *yqsm* (gloss: assures.3.sg.msc.imprf/promises.3.sg.msc.imprf; English: he swears), result in disagreement because in many cases even the context is ambiguous. In example 24, both interpretations of *it is unlikely to* and *it is not doable to* seem to be acceptable. Similarly, in example 25 يقسم *yqsm* can be interpreted as an a non-UC, meaning *promises*, or as a UC, meaning *assures*.

24.
- لا يمكن فهم كتاب #مرسي إلا بتأمل كتابين سرقات صغيرة وجنون الحكم.
- ***lA ymkn*** *fhm ktAb mrsy <lA bt>ml ktAbyn srqAt Sgyrp wjnwn AlHkm.*
- **not likely to/not doable to** understanding book #Morsi except by-contemplating books.dual robberies small and-mania the-ruling.
- **It is unlikely to /it is not doable to** understand Morsi's book without reading the other two books of *Small Robberies* and *Ruling Mania*.

25.
- عمرو أديب يقسم أن مصر لن تسقط.
- *Emr >dyb* ***yqsm >n*** *mSr ln tsqT.*
- Amr Adeeb **assures.3.sg.msc.imprf/promises.3.sg.msc.imprf that** Egypt not falls.3.sg.fm.imprf.
- Amr Adeeb **assures/promises that** Egypt will not fall.

UH-related disagreement is attributed mainly to long, syntactically complex clauses encoding holders. Syntactic complexity results from multiple coordinating phrases as in example 26, recursive descriptive relative clauses as in example 27, and apposition[9] as in example 28, among many other linguistic structures.

26.
- مبارك وطنطاوي والمجلس العسكري ومؤيدينهم واتباعهم واللي يتشدد لهم شايفين إن الشيعة أخطر من اليهود.
- *mbArk wTnTAwy wAlmjls AlEskry wm&ydynhm wAtbAEhm wAlly yt$dd lhm* ***$Ayfyn <n*** *Al$yEp >xTr mn Alyhwd.*
- Mubarak and-Tantawy and-the-council the-military and-supporters-their and-followers-their and-who supports.3.sg.msc.imprf-them **think.3.pl.imprf that** the-Shiites more.dangerous than the-Jews.
- Muabarak, Tantawy, the Military Council, their followers, and their supporters **think that** the Shiites are more dangerous than the Jews.

27.
- الناس اللي مش عاجبها كلامي واللي بتتريق عليا مقتنعين ان مرسي كان شغال في ناسا.
- *AlnAs Ally m$ EAjbhA klAmy wAlly bttryq ElyA* ***mqtnEyn An*** *mrsy kAn $gAl fy nAsA.*

[9] Apposition is a linguistic structure in which two noun phrases are placed side by side with one phrase serving to identify the other in a different way.

- the-people who not like.3.sg.fm.imprf talk-my and-who mock.3.sg.fm.imprf on-me **convinced.pl that** Morsi was.3.sg.msc.prf working.3.sg.msc.imprf in NASA.
- The people, who do not like my argument and are mocking me, **are convinced that** Morsi was working for NASA.

28.
- شادي حمدي مدير بركنغز في قطر: من غير المرجح أن يكون الإخوان المسلمون جزءا من المشهد السياسي مستقبلا.
- *$Ady Hmyd mdyr brwkngz fy qTr:* ***mn gyr AlmrjH >n*** *ykwn Al<xwAn Almslmwn jz’A mn Alm$hd AlsyAsy mstqblA.*
- Shady Hamdy manager Brookings in Qatar: **from not the-likely that** be the-Brotherhood the-Muslim part from the-scene the-political future.
- Shady Hamdy, the manager of Brookings Qatar: **it is unlikely that** the Muslim Brotherhood will be politically active in the future.

Scope-related disagreement is attributed first to scopes encliticized to their UCs. In 29, the scope starts at the 3rd person object pronoun attached to the UC افتكر *Aftkr* (gloss: thought.3.sg.msc.prf; English: it thought). For one annotator, those object pronouns are not part of the scope. For the other, they are part of the scope.

29.
- الجيش افتكرهم عيال سيس.
- *Aljy$* ***Aftkr****hm EyAl sys.*
- the-army **thought.3.sg.msc.prf**-them kids idiots.
- The army leaders **thought** they are some idiot kids.

Another reason for scope annotation disagreement is very long, syntactically complex clauses such as the ones in example 30.

30.

- أعتقد أنه لن يتم التوصل لاتفاق بين قوى المعارضة حتى لو مر ١٠٠ عام حتى لو مات كل المصريين.
- ***>Etqd >nh*** *ln ytm AltwSl lAtfAq byn qwY AlmEArDp HtY lw mr 100 EAm HtY lw mAt kl AlmSryyn.*
- **think.1.sg.imprf that** not reach.3.sg.imprf.passive to-agreement among power the-opposition even if passed.3.sg.msc.prf 100 year even if died.3.sg.msc.prf all the-Egyptians.
- I **think that** the opposition will not get to an agreement, even if they spend a 100 years trying, even if all Egyptians die.

2.3 Related Work

As I mentioned earlier in Chapter 1, there is a plethora of work on uncertainty annotation for English (Rubin, 2007; Szarvas et al., 2008; Saurí and Pustejovsky, 2009; Matsuyoshi et al., 2010; Farkas et al., 2010; Rubinstein et al., 2013; Wei et al., 2013), Japanese (Hendrickx et al., 2012), Chinese (Cui and Chi, 2013), Portuguese (Hendrickx et al., 2012; Avila and Mello, 2013), and Hungarian (Vincze, 2014). However, prior to my own preliminary work (Al-Sabbagh et al., 2014a) and the work I presented in this chapter, there are no Arabic uncertainty-annotated corpora, to the best of my knowledge. In this section, I start with a brief overview of some prior work on uncertainty annotation in other languages. Afterwards, I compare and contrast my own work.

Rubin (2007) annotated 80 English newspaper articles from the New York Times Service for explicitly-encoded certainty cues. She did not annotate holders or scopes. Reported kappa κ inter-annotator agreement rate is only 0.33.

Szarvas et al. (2008) developed the BioScope corpus that consists of biomedical abstracts and full papers. They annotated the corpus for negation and uncertainty cues and their scopes, but not their holders. The annotation process was carried out by two independent linguist annotators and a chief linguist - also responsible for setting up the annotation guidelines - who resolved cases of disagreement. The resulting corpus comprises more than 20K sentences, of which 10% contain at least one negation/uncertainty cue and its scope. No kappa κ or alpha α rates are reported.

Saurí and Pustejovsky (2009) annotated factuality as explicitly-encoded by epistemic and evidential modality triggers in 208 English documents from the TimeBank and the TimeML corpora. They achieved kappa κ inter-annotator agreement rates of 0.88 for cues, 0.95 for holders, and 0.81 for labeling cues as fact, counterfact, probable, not probable, possible, not certain, certain but unknown output, and unknown or uncommitted.

Farkas et al. (2010) created the WikiWeasel corpus that contains English Wikipedia paragraphs annotated for weasels. According to Farkas et al., a word is a weasel if it creates an impression that something important has been said, but what is really communicated is vague, misleading, evasive, or ambiguous. Weasel words do not give a neutral account of facts, but rather an opinion without any backup or source. Example weasels are *some people*, *possibly*, *might*, and *many people*, among others. The WikiWeasel corpus was used for the CoNLL 2010 shared task for detecting hedges.

Hendrickx et al. (2012) proposed an annotation scheme for Portuguese to annotate knowledge, belief, doubt, and possibility. For each concept, they annotated cues, holders, and scopes. In addition, they labelled each cue as either affirmative or negative. Based on 50 sentences, their annotation scheme achieved a kappa κ inter-annotator agreement rate of 0.85 for cues. With the same annotation scheme, Avila and Mello (2013) annotated 20 texts from the Brazilian Portuguese Spontaneous Speech corpus. No inter-annotator agreement rates are reported.

Rubinstein et al. (2013) proposed a linguistically-motivated annotation scheme to annotate epistemic beliefs, knowledge, and possibilities, among other concepts, in the MPQA corpus of English texts (Wiebe et al., 2005). Information labelled for each concept incorporates: (1) polarity (i.e. affirmative vs. negative), (2) propositional arguments (i.e. targets or scope spans), (3) sources (i.e. holders), (4) background (i.e. linguistic constituents that describe the circumstances and priorities that the claim is based on), and (5) degree indicators (i.e. linguistic constituents that indicate the degrees of possibilities). The reported alpha α inter-annotator reliability scores are 0.89 for cues and 0.65 for scopes. The same scheme was applied to the Chinese Penn Treebank by Cui and Chi (2013) who reported an inter-annotator agreement rate of 0.81 for cues.

Wei et al. (2013) annotated 4,743 English tweets so that each tweet is labelled as either uncertain or certain, based on the annotators' judgements about the author's intended meaning rather than the presence of uncertainty cues. For those tweets annotated as uncertain, sub-class labels are also required to label uncertainty as either epistemic (i.e. possible or probable) or hypothetical (i.e. condition, doxastic, dynamic, external, or question). The kappa κ coefficient indicating inter-annotator agreement is 0.907 for the certain/uncertain binary classification and 0.827 for the fine-grained annotation of uncertainty types.

Vincze (2014) manually annotated Hungarian texts for uncertainty from two domains: (1) 9,722 sentences from the Hungarian Wikipedia, and (2) 5,481 sentences from the Hungarian criminal news portal. She categorizes uncertainty cues into: epistemic, dynamic, doxastic, investigation, condition, weasel, hedge, and peacock. The corpus is only annotated for uncertainty cues. No inter-annotator agreement/reliability rates are reported.

This brief overview shows that there are subtle differences among the uncertainty-annotated corpora in terms of:

- **the annotated languages, genres, and domains**: most corpora are for the English language (Rubin, 2007; Szarvas et al., 2008; Saurí and Pustejovsky, 2009; Farkas et al., 2010; Rubinstein et al., 2013; Wei et al., 2013), except for a few annotation projects for Portuguese (Hendrickx et al., 2012; Avila and Mello, 2013), Chinese (Wei et al., 2013), and Hungarian (Vincze, 2014). The most widely covered domains and genres for uncertainty annotation are Wikipedia (Farkas et al., 2010; Vincze, 2014), biomedical texts (Szarvas et al., 2008; Vincze, 2014), and newswire texts (Rubin, 2007; Szarvas et al., 2008; Vincze, 2014). Only Wei et al. (2013) work on tweets.

- **the definition of uncertainty**: for some annotation projects uncertainty comprises beliefs, knowledge, hypotheses, and possibilities. Yet, for others, uncertainty also comprises conditionals, questions, investigations (Wei et al., 2013; Vincze, 2014). Some annotation projects annotate negation and uncertainty simultaneously, others label each independently.

- **the linguistic encoding of uncertainty**: some annotation projects annotate uncertainty at the token level looking for such cues as epistemic and evidential modality triggers, hedges, and/or weasels. Other projects annotate uncertainty at the sentence level where a sentence is labeled as (un)certain if it has at least one uncertainty cue.

- **the annotation targets**: some annotation schemes focus only on annotating uncertainty cues. Other schemes annotate cues, holders, and scopes. Some other schemes annotate more attributes for the uncertainty cues such as polarity (Rubinstein et al., 2013; Cui and Chi, 2013), tense, or intensity (Saurí and Pustejovsky, 2009).

- **the annotation procedure**: in some annotation projects, two or more annotators worked independently and then inter-annotator agreement/reliability was measured. For other projects, there is a chief annotator who resolves disagreements and thus no inter-annotator agreement/reliability rates are reported (Szarvas et al., 2008).

The subtle differences among uncertainty annotation projects make corpora of the same domain with intersecting data minimally comparable. Vincze et al. (2011) compared the negation and speculation annotations of the biomedical Genia Event (GE) and BioScope (BS) corpora. Although the two corpora intersect in 958 abstracts and 8,942 sentences, the agreement rate between them is only 0.48 because each corpus defines uncertainty cues and their scopes in a different way. In GE, uncertainty can be encoded by a verb, an adjective, or a noun. Yet, in BS uncertainty is linguistically encoded by the predicate and its arguments where the role of the predicate can be fulfilled by a verb, a noun, or an adjective. In GE, scopes include subjects, yet in BS, scopes are only complements. These differences motivated Szarvas et al. (2012) to propose a unified approach to annotate uncertainty.

My work is orthogonal to this extensive literature on uncertainty annotation. I defined uncertainty as hypotheses and speculations, following most of uncertainty annotation schemes. I annotated uncertainty cues, holders, and scopes, that are also considered in many previous annotation schemes. I hired two independent annotators to label my corpus and measured inter-annotator reliability rates so as to give more credibility for my newly developed Arabic NLP resource. I followed the unified annotation approach of Szarvas et al. (2012) and labeled all continuous and discontinuous cues and scopes, regardless of their part of speech or their grammatical function.

2.4 Conclusion

In this chapter, I presented my novel uncertainty-annotated corpus of Arabic tweets. The corpus is annotated for uncertainty cues, holders, and scopes. The development of this corpus provides the necessary resource to build a comprehensive uncertainty automatic analyzer for Arabic tweets, whose first machine learning model for uncertainty detection, starts in the next chapter.

CHAPTER 3

UNCERTAINTY DETECTION

3.1 Introduction

In Chapter 2, I built an adequately large corpus annotated for uncertainty cues, holders, and scopes in Arabic tweets. The main target of developing this corpus was to provide the necessary resource to train, test, and evaluate supervised machine learning models for Arabic uncertainty automatic analysis. With this corpus now in hand, I start in this chapter the first machine learning model in my comprehensive automatic uncertainty analyzer. The model is to identify and extract uncertainty linguistic cues.

The remainder of this chapter is organized as follows: Section 3.2 describes my approach to Arabic uncertainty detection, including features, experimental setup, results, and a detailed error analysis; Section 3.4 concludes the chapter and gives a future outlook.

3.2 Approach

3.2.1 Task Description

I define uncertainty detection as a token sequence labelling problem and apply Support Vector Machines (SVMs) as my machine learning method. I use the YamCha implementation[1] that has been used for multiple sequence labeling problems, especially in the literature of Arabic NLP, including Shaalan et al. (2009); Habash and Roth (2011); Alkuhlani and Habash (2011). SVMs and Conditional Random Fields (CRFs) have been both used in the literature of uncertainty automatic analysis. To the best of my knowledge, only Prabhakaran (2010) has compared the performance of the two machine learning methods in the context of English uncertainty detection to find out that CRFs marginally improve prediction accuracy. I keep the comparison between the two machine learning methods for my three uncertainty-related tasks for a future work.

For uncertainty detection, the classifier is trained to label each token as the beginning of an uncertainty cue (B-C), inside an uncertainty cue (I-C), or outside any uncertainty cues (O-C). With this BIO scheme, I manage to identify both unigram and multiword uncertainty cues as in Table 3.1.

3.2.2 Classification Features

I use a rich set of nine feature categories illustrated in Table 3.2.

Contextual Features (CFs) describe the lexical and morpho-syntactic contexts of each given token. The lexical context is the sequence of tokens around each given token; whereas the morpho-syntactic context is the sequence of Part-of-Speech (POS) tags. The morpho-syntactic CFs are extracted via MADAMIRA v1.0 Pasha et al. (2014), a comprehensive toolkit for Arabic morphological analysis, tokenization, and POS tagging.

Dialectal Features (DFs) identify the Arabic dialect of each given token and each given tweet, as to whether it is Modern Standard Arabic (MSA) or Egyptian Arabic (EA).

[1] http://chasenorg/ taku/software/yamcha

A Unigram UC

Arabic	Trans.	Gloss	English	BIO
أنا	*>nA*	I	I	O-C
أظن	***>Zn***	**think.1.sg.imprf**	**think**	**B-C**
مفيش	*mfy$*	not-there-not	there is no	O-C
فايدة	*fAydp*	benefit	benefit	O-C
من	*mn*	from	from	O-C
المقاطعة	*AlmqATEp*	the-boycott	the boycott	O-C

A Multiword UC

Arabic	Trans.	Literal	Gloss	BIO
من	***mn***	**from**	**It is**	**B-C**
المتوقع	***AlmtwqE***	**the-expected**	**expected**	**I-C**
تعيين	*tEyyn*	appointing	to appoint	O-C
العريان	*AlEryAn*	Aleryan	Aleryan	O-C
رئيسا	*r}ysA*	prime	as a Prime	O-C
للوزراء	*llwzrA'*	minister	Minister	O-C

Table 3.1: UCs represented in the BIO scheme and formatted based on YamCha requirements

DFs can be informative for uncertainty detection because some words can function as uncertainty cues in one Arabic variety but not the other. For example, شكلنا *$klnA* functions as an uncertainty cue only in EA where it means *it seems that*, but in MSA it is either a common noun encliticized to a possessive pronoun meaning *our look*, or a perfective verb conjugated for the 1st person plural meaning *we formed.* Likewise, the particle قد *qd* functions as an uncertainty cue only in MSA, in which it means either *indeed* if followed by a perfective verb, or *may* if followed by an imperfective verb. Yet, in EA قد *qd*

No	Feature	Description
Contextual Features (CFs)		
1	lexical	token sequences around each token
2	morpho-syntactic	POS sequences around each token
Dialectal Features (DFs)		
3	token-dialect	the Arabic dialect of each token
4	tweet-dialect	the Arabic dialect of each tweet
Lexicon Feature (LFs)		
5	token-in-lexicons	the presence/absence of each token in the Arabic uncertainty lexicons
Semantic Features (SemFs)		
6	gender	the gender of each token, if applicable
7	number	the number of each token, if applicable
8	person	the person of each token, if applicable
Syntactic Features (SynFs)		
9	base-phrase-type	the type of the base phrase of which each token is a part
10	position-in-base-phrase	the position of each token within its base phase
11	syntactic-dependencies	syntactic dependencies of each token
Twitter Features (TFs)		
12	tweet-length	the number of tokens per tweet
13	token-position	the position of each token within its tweet
14	hashtag-count	the number of hashtags within each tweet, if any
15	URL-count	the number of URLs within each tweet, if any

Table 3.2: Classification features for uncertainty detection

is only a comparative particle meaning *as ... as*. Furthermore, DFs can be informative for uncertainty attribution: we assume that users are more likely to use their native local Arabic dialect, i.e., EA, when they are tweeting casually about their own speculations and hypotheses. DFs are extracted via the Arabic dialect identifier, AIDA (Elfardy et al., 2014).

Lexicon Feature (LF) is a binary feature: if a given token is in the Arabic uncertainty lexicons, the feature value is set to *true*; otherwise to *false*. For this feature, I built two lexicons: the first is a manually-compiled lexicon of 3,289 unigram UCs (Al-Sabbagh et al., 2013); and the second is an automatically-generated lexicon of 4,795 multiword UCs (Al-Sabbagh et al., 2014b).

The unigram lexicon is part of a larger lexicon that comprises unigrams expressing different types of modality: epistemic, evidential, obligative, permissive, commissive, abilitive, and volitive (Al-Sabbagh et al., 2013). Epistemic modality is defined as the speaker's judgment about the factual status of the proposition, whereas evidential modality is restricted to hearsay and sensory expressions such as قال *qAl* (gloss: said.3.sg.msc.prf; English: he said) and سمع *smE* (gloss: heard.3.sg.msc.prf; English: he heard), respectively. As a result, I only use epistemic modality expressions from that lexicon as they are the closest to my definition of uncertainty in this research project. The entire lexicon is manually generated as I compiled unigram modality expressions from several theoretical studies, including (Mitchell and Al-Hassan, 1994; Brustad, 2000; Badawi et al., 2004), and then I manually generated the morphological inflections and derivations of each compiled expression. Furthermore, I manually added the English translation for each entry. The epistemic modality portion of this lexicon is 3,289 unigrams.

Arabic	Transliteration	Gloss	English
أظن	*>Zn*	think.1.sg.imprf	I think
نظن	*nZn*	think.1.pl.imprf	we think
يظن	*yZn*	thinks.3.sg.msc.imprf	he thinks
عارف	*EArf*	know.1.sg.imprf/knows.3.sg.msc.imprf	I/he know(s)
عارفة	*EArfp*	know.1.sg.imprf/knows.3.sg.fm.imprf	I/she know(s)
عارفين	*EArfyn*	know.1.pl.imprf/know.2.pl.imprf/know.3.pl.imprf	we/you/they know

Table 3.3: An excerpt from the lexicon of unigram UCs

The second lexicon is also part of a larger project to identify multiword modality expressions Al-Sabbagh et al. (2014b), including multiword expressions for epistemic, evidential, obligative, permissive, commissive, abilitive, and volitive modality. I also used the epistemic modality multiword expressions only for my research project here. This is because epistemic modality was also defined as the speaker's judgement about the factual status of the proposition; whereas evidential modality is restricted to hearsay and sensory expressions. The number of multiword expressions that denote epistemic modality, and hence, uncertainty is 4,795 expressions. The lexicon is automatically generated using a *k*-means clustering algorithm and a large number of morphological, syntactic, and lexical features. For more details on both lexicons, I refer the reader to Al-Sabbagh et al. (2013) and Al-Sabbagh et al. (2014b). Excerpts from both lexicons are illustrated in Tables 3.3 and 3.4, respectively.

Semantic Features (SemFs) describe the gender, number, and person features for each token, if applicable. SemFs are especially informative for uncertainty attribution. According to Arabic syntax, if the cue is a verb, a present participle, a noun, or an adjec-

Arabic	**Transliteration**	**Gloss**	**English**
من المؤكد أن	*mn Alm&kd >n*	from the-sure that	it is sure that
لدي شعور بأن	*ldy $Ewr b>n*	have-me feeling with-that	I have a feeling that
على قناعة بـ	*ElY qnAEp b>n*	on conviction with	(be) absolutely sure that
كلي يقين في أن	*kly yqyn fy >n*	all-me confidence in that	I am all confident that

Table 3.4: An excerpt from the lexicon of multiword UCs

tive, I have to expect its holder to have the same gender, number, and person features. To extract SemFs, I rely on a few resources:

- The ATB tagset: MADAMIRA v1.0 (Pasha et al., 2014) uses the Penn Arabic TreeBank (ATB) tagset (Maamouri et al., 2009), which explicitly encodes gender, number, and person if and only if they are morphologically marked by such affixes as: the feminine plural suffix ات *At* as in بنات *bnAt* (gloss: girls.pl.fm; English: girls), the feminine singular suffix ة *p* as in ابنة *Abnp* (gloss: daughter.sg.fm; English: a daughter), the 3rd person imperfective prefix ي *y* as in يعتقد *yEtqd* (gloss: thinks.3.sg.msc.imprf; English: he thinks), and the 1st person plural prefix ن *n* (n) as in نعتقد *nEtqd* (gloss: think.1.pl.imprf; English: we think), among many other.

- Since gender, number, and person are not always morphologically represented, I also use the Arabic lexicon of semantic features from Elghamry et al. (2008) that comprises 30,000 entries labeled for gender and number.

- I also use the database from Alkuhlani and Habash (2011) that comprises the words of the Penn Arabic TreeBank labeled for gender and number, among other semantic features.

Syntactic Features (SynFs) comprise two types of features: (1) **shallow parsing features** that describe the type of the base phrase of which each token is a part, and the position of each token within its base phase; and (2) **dependency parsing features** that describe the syntactic dependencies among the parts of complex clauses. Many cues, holders, and scopes are base phrases. Example base phrase cues are: the prepositional phrase على الأغلب *ElY Al>glb* (gloss: on the-most; English: most probably) and the adverbial phase ربما *rbmA* (gloss: maybe; English: maybe). Likewise, holders and scopes can be base phrases such as the base noun phrase holder الرئيس *Alr}ys* (gloss: the-president; English: the president), and the base deverbal noun scope نجاح *njAH* (gloss: success; English: success) in examples 1 and 2, respectively.

1.
   - يعتقد الرئيس أن الخونة يسأكلوننا.
   - ***yEtqd*** *Alr}ys* *>**n*** *Alxwnp sy>klwnnA*.
   - **thinks-3.sg.msc.imprf** the-president **that** the-traitors will-eat.3.pl.imprf-us.
   - The president **thinks that** we will be defeated by the traitors.

2.
   - الشعب غير واثق من نجاح مرسي
   - *Al$Eb* ***gyr wAvq mn*** *njAH mrsy*
   - the-people **not sure.sg.msc from** success Morsi
   - The people **are not sure that** Moris can succeed.

Complex cues, holders, and scopes are not uncommon, however, especially that the annotation guidelines of the corpus I use are based on Szarvas et al. (2008)'s maximal length principle, according to which the marked text segments for holders and scope must include all related complements and adjuncts. Consequently, I decide to use both base phrase and syntactic dependency information. Shallow parsing features are extracted via the Arabic shallow parser of AMIRA v2.0 Diab (2009). Dependency parsing features are extracted via the CATiB dependency parser (Marton et al., 2013).

Twitter Features (TFs) describe for each tweet (1) its length (i.e. number of tokens), (2) the number of hashtags, if any, (3) the number of URLs, if any, and (4) the position of each token in the tweet. Twitter-based features have been found useful for English uncertainty detection Wei et al. (2013).

3.2.3 Experimental Setup and Results

I use the same experimental setup for each task in my comprehensive automatic uncertainty analyzer. To find my optimal machine learning model per uncertainty-related task, I implement a 10-fold cross validation method in which the whole corpus is partitioned into 10 disjoint segments: for each fold I train on 9 segments and test on the 10th. All reported accuracy, precision, recall, and F_1 score rates are averaged across the 10 folds. For each task, I run my experiments to find the optimal:

- **Feature category combination**, using a greedy algorithm. For the first round of the algorithm, I start by evaluating each feature category on its own, and then I select the highest performing feature category, compared to the baseline. For the second round of the algorithm, I combine the best category from the first round with each of the rest feature categories, making 2-feature-category combinations; and then I select the highest performing 2-feature-category combination. For the third round, I use the best combination from the second round, and combine it with one feature category at a time, forming 3-feature-category combinations; and then I select the highest performing 3-feature-category combination. I continue the algorithm until I reach the largest best combination of feature categories.

- **Linear context width**, which is the window of tokens whose features are considered. For instance, a linear context width of ±2 means that the feature vector for any given token includes, in addition to its own features, those of the 2 tokens before and after it as well as the predictions of the 2 tokens before it.
- **Polynomial degree**, starting with 2, the default polynomial degree of YamCha SVMs implementation.
- **Parsing direction**: forward (left to right) vs. backward (right to left).
- **Multiclass method**: one-against-the-rest vs. pairwise.

As a baseline, I use a lexicon look-up model given that lexicons of Arabic uncertainty cues do exist. This baseline is the same as the lexicon feature number 5 from Table 3.2. A similar lexicon look-up baseline has been used for Hungarian uncertainty detection (Vincze, 2014).

My experiments show that one-against-the-rest multiclass classification in a forward parsing direction (i.e. left to right) is the best configuration with the default YamCha polynomial kernel degree of 2. The optimal linear context width is found to be ±4.

3.2.4 Discussion

According to Table 3.5, my greedy algorithm for feature selection finds that the best stand-alone feature category, compared to the baseline, is the CFs category. This is ex-

No	Feature Category Combinations	Accuracy	Precision	Recall	F_1
0	Baseline	0.451	0.428	0.538	0.477
1	CFs	0.778	0.748	0.723	0.735
2	CFs+SynFs	0.835	0.799	0.782	0.790
3	CFs+SynFs+LF	0.856	0.832	0.791	0.811
4	CFs+SynFs+LF+SemFs	0.888	0.848	0.793	0.819
5	CFs+SynFs+LF+SemFs+DFs*	0.904	0.854	0.813	0.838
6	CFs+SynFs+LF+SemFs+DFs+TFs	0.909	0.849	0.830	0.839

Table 3.5: Results for uncertainty detection with the best feature category combination marked with an asterisk

pected. About 78.65% of the cues in my corpus are multiword expressions that consist of a head verb/noun/adjective and (1) a complementizer as in يعرف إن *yErf <n* (gloss: knows.3.sg.msc.imprf that; English: he knows that), (2) a preposition as in واثق في *wAvq fy* (gloss: sure.sg.msc in; English: sure that), or (3) a preposition and a complementizer as in يؤمن بأن *y&mn b>n* (gloss: believes.3.sg.msc.imprf in-that; English: he believes that). CFs contribute to identifying such cue-distinguishing subcategorization frames starting with complementizer and prepositions.

In the second round of the feature selection algorithm, the optimal 2-feature category combination comprises the CFs and SynFs, with an F_1 increase of 0.055. compared to the first round of the algorithm. Out of 13,620 multiword cues in my corpus, 10,544 do not have any linguistic constituents in-between their parts, 1,875 have one in-between linguistic constituent, 710 have two in-between constituents, 260 have three in-between constituents, and the rest have four or more in-between constituents. As a result, SynFs significantly improve performance for uncertainty detection, by detecting base phrase

UCs and relating the different non-adjacent parts to the heads of the multiword UCs.

SemFs introduce a small precision increase, which is statistically significant according to my paired *t*-test on the 10-fold cross validation runs (*p*-value = 0.01228). Similarly, DFs increase the recall rate with a *p*-value of 0.0093, that is also statistically significant.

In the last round of my greedy feature selection algorithm, the highest performing feature category combination includes the TFs. Yet, the difference between the F_1 scores of combinations numbers 5 and 6 is only 0.001, which is not statistically significant (*p*-value = 0.768). Consequently, I stop my search for the optimal combination of feature categories and consider combination number 5 as my best for uncertainty detection.

3.2.5 Error Analysis

In the output of my uncertainty detector, I identify five main error triggers, arranged below from the most to the least frequent. The first error trigger is tokens that occur in the same lexical and morpho-syntactic context, whether they convey uncertainty or not. Although افتكرت *Aftkr* in example 3 denotes uncertainty (gloss: thought.1.sg.prf; English: I thought) and in 4 it does not (gloss: remembered.1.sg.prf; English: I remembered), in both examples it comes at a tweet-initial position followed by a complementizer. Similarly, علمت *Elmt* in example 5 is a UC (gloss: realized.3.sg.fm.imprf; English: they realized) and in example 6 is not a UC (gloss: taught.3.sg.fm.prf; English: it taught). Yet, in both examples it occurs in the same lexical and morpho-syntactic context.

3. - افتكرت ان لما الناس حتشوف التعذيب حتثور.
   - ***Aftkrt An*** *lmA AlnAs Ht$wf AltEzyb Htvwr.*

- **thought.1.sg.prf that** when the-people will-witness.3.sg.fm.imprf the-torture will-rebel.3.sg.fm.imprf.
- I **thought that** the people will rebel when they witness the torture.

4. - افتكرت ان الناس شافت بنت بتتسحل وسكتت.
   - *Aftkrt An AlnAs $Aft bnt bttsHl wsktt.*
   - remembered.1.sg.prf that the-people witnessed.3.sg.fm.prf girl.sg.fm tortured.3.sg .fm.prf.passive and-remained.silent.3.sg.fm.prf.
   - I remembered that the people witnessed a girl being tortured and remained silent.

5. - علمت الناس انه لا سبيل سوى الديموقراطية.
   - ***Elmt*** *AlnAs* ***Anh*** *lA sbyl swY AlDymwqrATyp*.
   - **realized.3.sg.fm.prf** the-people **that** no way but the-democracy.
   - The people **realized that** democracy is the only option.

6. - الديكتاتورية علمت الناس ان من خاف سلم .
   - *AldyqtAtwryp Elmt AlnAs An mn xAf slm.*
   - the-tyranny taught.3.sg.fm.prf the-people that who scared.3.sg.msc.prf safe.sg.msc.
   - Tyranny taught the people that it is better afraid than sorry.

The second error trigger is highly-biased tokens such as لازم *lAzm* .In 97.4% of its occurrences, it denotes obligation as in example 7, and in the rest it denotes uncertainty as in example 8.

7. - لازم الشعب يفوق قبل فوات الأوان.
   - *lAzm Al$Eb yfwq qbl fwAt Al>wAn.*
   - must the-people wake-up.3.sg.msc.imprf before missing the-opportunity.

- The people must wake up before it is too late.

8. - لازم مرسي خايف من الصوابع.
   - ***lAzm*** *mrsy xAyf mn AlSwAbE*.
   - **must** Morsi afraid.sg.msc of the-fingers.
   - **It must be that** Morsi is afraid of conspiracies.

The third error trigger is discontinuous multiword UCs with very long in-between linguistic constituents. In example 9, the complementizer أن >*n* (gloss: that; English: that) is nine tokens apart to the right of its head verb استبعدت *AstbEdt* (gloss: excluded.3.sg.fm.prf; English: she excluded the possibility), because the noun phrase that represents the holder falls in-between.

9. - استبعدت داليا زيادة المدير التنفيذي لمركز ابن خلدون للدراسات الأنمائية أن يكون روبرت فورد هو السفير الأمريكي.
   - ***AstbEdt*** *dAlyA zyAdp Almdyr Altnfyzy lmrkz Abn xldwn lldrAsAt Al<nmA}yp* >***n*** *ykwn rwbrt fwrd hw Alsfyr Al>mryky*.
   - **excluded.3.sg.fm.prf** Mai Zeyada the-executive the-manager for-center Ibn Khaldwn for-the-studies the-developmental **that** be Robert Ford the-ambassador the-American.
   - Mai Zeyada, the executive manager of Ibn Khaldwn Center for Developmental Studies, **excluded the possibility that** Robert Ford is the (coming) American ambassador.

The fourth error trigger is UCs with incomplete subcategorization frames. As we mentioned earlier, one reason that CFs and SynFs yield high results for uncertainty detection is that they contribute to identifying cue-distinguishing subcategorization frames starting with complementizers and/or prepositions. Sometimes, due to stylistic preferences, such complementizers and prepositions are removed. Hence, cues miss one key identifying feature as in example 10.

10.
   - متهيالي الناس مش حتسكت المرة دي.
   - ***mthyAly*** *AlnAs m$ Htskt Almrp dy.*
   - **think.1.sg.imprf** the-people not will-remain.silent.3.sg.fm.imprf the-time this.
   - I **think** people will not let it go this time.

Finally, tokenization and POS tagging errors contribute to the uncertainty detection errors, especially that the current available version of MADAMIRA v1.0 does not fully support Arabic dialects that make a good portion of my corpus.

3.3 Related Work

Approaches to automatic uncertainty detection are either token or sentence level. Token level approaches define uncertainty detection as a token sequence labeling problem, where classifiers label each token as Beginning-UC, Inside-UC, or Outside-UC. Sentence level approaches train classifiers to label each sentence as uncertain if it has at least one UC, or as certain, otherwise.

Token level approaches may have some advantages over sentence level approaches. First, they can process unigram and multiword UCs in one fell-swoop. Approaches that define uncertainty detection at the sentence level have to build post processing models to process multiword UCs such as Täckström et al. (2010). Second, they facilitate further processing to identify and extract uncertainty holders and scopes by identifying the text segments that represent the UCs. As a result, I use token level approaches for my uncertainty detector. In this section, I present a brief overview of token level approaches to uncertainty detection; and compare and contrast my work to others'.

Tang et al. (2010) participated in the CoNLL Shared Task 2010 (Farkas et al., 2010) and used Conditional Random Fields (CRFs) to detect hedges in biological and Wikipedia corpora. They used a rich linguistic feature set with some features as *n*-grams, POS sequences, and token position. They reported a performance F_1 score of 0.864 for the biological corpus and of 0.551 for the Wikipedia corpus.

Prabhakaran (2010) compared the performance of SVMs and CRFs to detect English UCs in the same CoNLL Shared Task 2010. He achieved an F_1 score of 0.429 with a rich feature set of three categories: lexical features, word list features, and syntactic features. He found that syntactic features marginally improve performance. This is because the syntactic patterns that proved helpful for this task were fairly local. So, probably exploring shallow syntactic features instead of deep syntactic features or using custom made lexicons could also improve performance. As for the comparison between SVMs and CRFs, he found that CRFs marginally improved the prediction accuracy while substantially improved the speed.

Zhao et al. (2010) combined CRFs with POS information and a lexicon of English hedges and weasels for the same CoNLL Shared Task 2010. Although their system yielded an F_1 score of 0.753 for the biomedical genre, it only got an F_1 score of 0.312 for the Wikipedia genre. Yet, they assured that a lexicon of hedges and weasels is important for cross-domain applicability.

Szarvas et al. (2012) addressed uncertainty cue detection in a multi-domain setting, using surface level, part-of-speech, and chunk-level features, and CRFs. They found that cue words can be accurately detected in texts with various topics and stylistic properties. Their results suggest that simple cross training can be employed and it achieves a reasonable performance (60 to 70% cue-level F_1 score) when no annotated data is at hand for a new domain. When some annotated data is available, domain adaptation techniques are the best choice.

Vincze (2014) applied token level sequence labeling methods and a rich linguistic feature set to detect UCs in Hungarian, a morphologically-rich language. She got 0.396 and 0.449 for macro and micro F_1 scores, respectively. Similar to Vincze, I work on a morphologically-rich language, use a lexicon look-up baseline, apply sequence labeling methods, and rely on a rich linguistic feature set. Unlike Vincze, however, I simplify uncertainty detection so that the classifier makes the three-way decision of whether a given token is B-UC, I-UC, or O-UC, regardless of the UC type, whether that is an epistemic modality trigger, a hedge, or a weasel. This simplification is a main reason for the higher performance that I obtain. Neither Vincze nor Wei et al. (2013), who use the same fine-grained classification of UC types, discusses its practical implications for NLP applications. Hence, my simplification might be a better approach for uncertainty detection. Other differences between my work and Vincze's are that: (1) I use SVMs but she uses CRFs; (2) she works on Wikipedia and newswire texts; and (3) unlike her I use genre-specific features.

Although Wei et al. (2013) define uncertainty detection at the sentence level, I am interested in mentioning their work because it is the only study that works on the linguistic genre of tweets in the context of uncertainty detection. With content-based, Twitter-based, and user-based features, they obtained an F_1 score of 0.822. Similar to Wei et al. (2013), I work on the linguistic genre of tweets and use genre-specific features. However, they define uncertainty detection at the sentence level, where a sentence is labeled as uncertain if it has at least one UC. I think token level uncertainty detection might be more convenient to detect unigram and multiword UCs in one fell-swoop, instead of developing post processing modules to process them as in Täckström et al. (2010). My feature set is more linguistically elaborate compared to theirs although they use more genre-specific features than me. Their genre-specific features include the counts for the URLs, hashtags, retweets, and replies per tweet as well as the counts for the followers, lists, friends, tweets, and favorites per Twitter user. In my opinion, to use linguistic features pertinent to uncertainty itself, rather than to the genre which is being studied, is to increase the applicability of my models to other genres. As a result, I did not want to rely heavily on genre-specific features as in Wei et al. (2013).

3.4 Conclusion

In this chapter, I presented the first part of my comprehensive Arabic automatic uncertainty analyzer, namely the uncertainty detector to identify and extract UCs in each given tweet. Once UCs are identified, I can further process them to identify and extract their holders and scopes as I do in the next two chapters, respectively. For uncertainty detection, I used a rich feature set of linguistic and non-linguistic features. With SVM sequence labeling and based on the average rates of 10-fold cross validation, I achieved an F_1 score of 0.838. The result is promising given the reported results on uncertainty detection for other languages.

CHAPTER 4

UNCERTAINTY ATTRIBUTION

4.1 Introduction

Once UCs are identified, the second task in my comprehensive automatic uncertainty analyzer is to ascribe each identified cue, one cue at a time, to its holder. In this chapter, I present my machine learning model for uncertainty attribution; which is an understudied task in uncertainty NLP research: some researchers such as Baker et al. (2012) overlook holder annotation, identification, and extraction; while others either set text writers as the default holders (Diab et al., 2009), or use a predefined set of prototypical holders (Wiegand and Klakow, 2011a).

Default and prototypical holders are unlikely to work for uncertainty attribution in the highly-interactive linguistic genre of tweets in which uncertainty holders are not necessarily the users who posted the uncertainty-laden tweets; instead, they can be some other holders assumed or cited by the posting users. Furthermore, ignoring uncertainty attribution affects the performance of NLP applications that are more concerned with uncertainty holders than with uncertainty cues. Examples of those NLP applications are: credibility analyzers that detect disinformers who endorse rumors and further spread them (Castillo et al., 2011; Soni et al., 2014), and topical expertise finders that select trustful holders with experience of specific topics (Wagner et al., 2012).

The rest of this chapter is organized as follows: Section 4.2 describes my approach to Arabic uncertainty attribution, including task description, classification features, experimental setup, experimental results, a discussion of the results, and an error analysis; and Section 4.3 compares and contrasts my approach to closely-related approaches.

4.2 Approach

4.2.1 Task Description

Similar to uncertainty detection, uncertainty attribution is defined as a token sequence labeling problem, in which the classifier predicts for each given token whether it is the beginning of a holder (B-H), inside a holder (I-H), or outside any holders (O-H). As I mentioned in Section 2.2.2.2, there are three main types of holders in my corpus. For holders encoded in the morphological inflections or the semantics of their UCs, I place the BIO-H labels on the UCs themselves as Table 4.1 shows. According to my corpus statistics from Section 2.2.3, 7,992 out of 17,317 holders in my corpus are encoded in their UCs.

4.2.2 Classification Features

The classification features for uncertainty attribution comprise all the feature categories used for uncertainty detection in Table 3.2, in addition to two new feature categories described in Table 4.2: Pragmatic Features (PFs) and Uncertainty Cue Features (UCFs).

Arabic	Trans.	Gloss	English	BIO
مرسي	*mrsy*	Morsi	Morsi	B-H
فاكر	***fAkr***	**thinks.3.sg.msc.imprf**	**thinks**	**O-H**
نفسه	*nfsh*	himself	he is	O-H
إله	*≤lh*	god	a god	O-H

مظنش	***mZn$***	**not-think.1.sg.imprf-not**	I **do not think**	**B-H**
أن	***≥n***	**that**	**that**	**I-H**
عمر	*Emr*	Omar	Omar	O-H
سليمان	*slymAn*	Suleiman	Suleiman	O-H
قادر	*qAdr*	capable.sg.msc	is capable	O-H
على	*ElY*	on	of	O-H
خوض	*xwD*	fighting	fighting	O-H
معركة	*mErkp*	battle	a battle	O-H

Table 4.1: Uncertainty holders represented in the BIO scheme and formatted based on YamCha requirements

Pragmatic Features (PFs) comprise two features: (1) a binary feature to determine whether there are linguistic markers for (in)direct reported speech, including quotation markers and the reported speech verbs of قال *qAl* (gloss: said.3.sg.msc.prf; English: he said), زعم *zEm* (gloss: claimed.3.sg.msc.prf; English: he claimed), صرح *SrH* (gloss: declared.3.sg.msc.prf; English: he declared), أعلن *>Eln* (gloss: announced.3.sg.msc.prf; English: he announced), عبر *Ebr* (gloss: expressed.3.sg.msc.prf; English: he expressed), أعرب

>Erb (gloss: stated.3.sg.msc.prf; English: he stated), conjugated for different persons, genders, numbers, and aspects; and (2) a binary feature to locate each token as either occurring before or after the linguistic markers of (in)direct reported speech, if there are any. Based on my corpus observations, when a direct quote has UCs, UHs come before the colon (:), which is the typical punctuation marker used with direct reported speech as in example 1. In contrast, when an indirect quote has UCs, the UHs typically come after the linguistic marker of the indirect reported speech as in example 2.

1.
   - أمير قطر: أؤمن أن الوطن العربي جسد واحد وأوصيكم بالثبات على الحق.
   - *>myr qTr*: *>**&mn** >**n** AlwTn AlErby jsd wAHd w>wSykm bAlvbAt ElY AlHq.*
   - prince Qatar: **believe.1.sg.imprf that** the-world the-Arab body one and-ask.1.sg .imprf-you.pl to-the-sticking on the-right.
   - The prince of Qatar: I **believe that** the Arab world is a unity and I ask you to stick to what is right.

2.
   - أعلن المجلس الوطني الانتقالي الليبي أنه يتوقع سقوط سرت بالكامل.
   - *>Eln Almjls AlwTny AlAntqAly Allyby >nh **ytwqE** sqwT srt bAlkAml.*
   - declared.3.sg.msc.prf the-council the-national the-transitional the-Libyan that-it **expects.3.sg.msc.imprf** collapse Sert by-the-full.
   - The Libyan National Transitional Council declared that it **expects** the full collapse of Sert.

Uncertainty Cue Features (UCFs) are extracted from the output of the machine learning model for uncertainty detection from Chapter 3 and are used for the next two tasks in the pipeline, namely uncertainty attribution and scope extraction. This is the key point of using a unified framework for uncertainty automatic analysis: the predictions of

No	Feature	Description
Pragmatic Features (PFs)		
1	reported-speech	the presence/absence of (in)direct reported speech linguistic markers
2	token-location	the location of each token as to whether it comes before or after the (in)direct reported speech linguistic markers, if any
Uncertainty Cue Features (UCFs)		
3	cue	text segments representing identified cues in each tweet
4	cue-position	the position of each identified cue in its tweet
5	cue-length	whether each identified cue is a unigram or a multiword expression
6	cue-location	whether each token comes before or after the identified cue in each tweet
7	cue-distance	the distance between each token and the identified cue in each tweet

Table 4.2: Two more classification features for uncertainty attribution

one machine learning model inform the other models. For each identified UC, I describe the following UCFs:

- **Cue:** the text segment representing the cue.
- **Cue-Length:** whether the cue is a unigram or a multiword expression.
- **Cue-Position:** the position of the cue in the tweet. Typically, cues at tweet-initial positions have their holders encoded in their morphological inflections for person.

- **Cue-Location:** whether each token comes before or after the identified cue.
- **Cue-Distance:** the distance between each token and the identified cue, defined as a numeric value.

4.2.3 Experimental Setup and Results

I use the same experimental setup from Section 3.2.3: I split the corpus into 10 disjoint segments and implement a 10-fold cross validation method, in which for each fold I train on 9 segments and test on the 10th. All reported accuracy, precision, recall, and F_1 results are averaged across the 10 folds. I use the same greedy algorithm to select the best feature category combination and also examine the optimal linear context width, parsing direction, multiclass method, and polynomial degree.

As a baseline, I use a simple bag-of-words model, based on token sequences around each given token. This baseline model is the same as the lexical contextual feature number 1 in Table 3.2.

Similar to uncertainty detection, one-against-the-rest multiclass classification in a forward parsing direction (i.e. left to right) is the best configuration with the default YamCha polynomial kernel degree of 2. The optimal linear context width is found to be -10 and +4. It is expected for uncertainty attribution to need a larger linear context width to find the holder of each identified cue, one cue at a time, given the following three facts about Arabic syntax: (1) the flexible word order of Arabic accepts UHs to precede or follow their UCs; (2) long dependencies between UCs and their UHs occur more frequently when UHs precede their UCs as in example 3; that is why the optimal left linear context width for uncertainty attribution is as large as -10; and (3) UHs tend to closely follow their UCs

No	Feature Category Combinations	Accuracy	Precision	Recall	F_1
0	Baseline	0.468	0.413	0.489	0.448
1	CFs	0.664	0.698	0.708	0.703
2	CFs+SynFs	0.699	0.733	0.718	0.725
3	CFs+SynFs+UCFs	0.727	0.761	0.742	0.751
4	CFs+SynFs+UCFs+PFs	0.736	0.770	0.747	0.758
5	CFs+SynFs+UCFs+PFs+SemFs	0.742	0.776	0.750	0.763
6	CFs+SynFs+UCFs+PFs+SemFs+TFs*	0.750	0.784	0.762	0.773
7	CFs+SynFs+UCFs+PFs+SemFs+TFs+DFs	0.746	0.780	0.768	0.774

Table 4.3: Results for uncertainty attribution with the best feature category combination marked with an asterisk

when the UHs follow their UCs as in example 4; as a result, the best right linear context width is only +4.

3. - جلال أمين: الدولة البوليسية ليست دولة قوية بل دولة فاسدة في رأيي.
   - *jlAl >myn: Aldwlp Albwlysyp lyst dwlp qwyp bl dwlp fAsdp* ***fy r>yy***.
   - Galal Ameen: the-state the-police not state strong but state corrupt **in opinion-my**.
   - Galal Ameen: the police state is not a strong state, but a corrupt one, **in my opinion**.

4. - يحسب الناس أن يتركوا دون عقاب.
   - ***yHsb*** *AlnAs* ***>n*** *ytrkwA dwn EqAb*.
   - **think.3.sg.msc.imprf** the-people **that** left.3.pl.imprf.passive without punishment.
   - The people **think that** they will not be punished.

Similar to uncertainty detection, CFs and SynFs win the second round of the greedy feature selection algorithm. As per expectation, morpho-syntactic and syntactic features abstract away from the surface tokens, that are highly variable given Arabic rich morphology; and, hence they can capture phrase and clause structures encoding holders more successfully.

UCFs significantly improve performance with an F_1 score increase of 0.026. One main advantage of using UCFs is reducing the number of tokens to be considered for uncertainty attribution. As we mentioned earlier, only 7,992 holders out of 17,317 are encoded in the morphological inflections of their UCs. This entails that in the majority of cases a token that has been labeled as B-C or I-C is unlikely to be considered for uncertainty attribution. This elimination process of noisy tokens is one main advantage of my proposed unified framework, in which the predictions of one machine learning model informs the predictions of the next machine learning model in the pipeline.

PFs make it to the fourth round of the feature selection greedy algorithm. As I mentioned earlier, I have noticed that (in)direct reported speech is very systematically structured in my corpus: for indirect reported speech, UHs typically come after the reported speech linguistic markers; and for direct reported speech, UHs tend to come before the reported speech linguistic markers.

SemFs and TFs introduce small, yet statistically significant, improvements with paired t-test p-values of 0.0136 and 0.0345, respectively. Yet, DFs do not yield any significant improvements as I compare the outputs of the sixth and seventh rounds of my feature selection greedy algorithm; paired t-test p-value = 0.837.

4.2.4 Error Analysis

Three main factors contribute to uncertainty attribution errors. The first and the most frequent is long, syntactically complex clauses encoding UHs as in example 5.

5.
- استبعدت داليا زيادة المدير التنفيذي لمركز ابن خلدون للدراسات الأنمائية أن يكون روبرت فورد هو السفير الأمريكي.
- ***AstbEdt*** *dAlyA zyAdp Almdyr Altnfyzy lmrkz Abn xldwn lldrAsAt Al<nmA}yp* ***>n*** *ykwn rwbrt fwrd hw Alsfyr Al>mryky.*
- **excluded.3.sg.fm.prf** Mai Zeyada the-executive the-manager for-center Ibn Khaldwn for-the-studies the-developmental **that** be Robert Ford the-ambassador the-American.
- Mai Zeyada, the executive manager of Ibn Khaldwn Center for Developmental Studies, **excluded the possibility that** Robert Ford is the (coming) American ambassador.

The second factor contributing to uncertainty attribution errors is holders inserted in-between the boundaries of multiword expressions as example 6 below.

6.
- أكد الناشط الحقوقي نجاد البرعي أنه لا يوجد سبب واضح للهجوم على المراكز.
- *>**kd** AlnA$T AlHqwqy njAd AlbrEy >**nh** lA ywjd sbb wADH llhjwm ElY AlmrAkz.*
- **assured.3.sg.msc.prf** the-activist the-humanitarian Nijad Alborei **that** no exists.3.sg.msc.imprf reason clear to-attack on the-headquarters.
- Nijad Alborei, the humanitarian activist, **assured that** there is no clear reason for attacking the headquarters.

The third factor is long dependencies between UCs and their holders, even with base phrase holders. In example 7, the holder is the base noun phrase ماما *mAmA* (gloss: mum;

English: mum); yet it is six tokens apart from its UC due to the in-between verb phrase.

7.
   - ماما لسه مكلماني من عند السفارة بتقول مفيش احتمال ان الناس تدخل العمارة.
   - *mAmA lsh mklmAny mn End AlsfArp btqwl **mfy$ AHtmAl An** AlnAs tdxl AlEmArp.*
   - Mum just called.3.sg.fm.prf-me of from the-embassy saying.3.sg.fm.imprf **not-is-not possibility that** the-people break.into.3.sg.fm.imprf the-building.
   - Mum has just called me from the embassy, saying **it is unlikely that** the people will break in.

4.3 Related Work

Although there is no prior work specifically on uncertainty attribution, attribution has been extensively considered in the context of opinion mining to ascribe opinions to their holders.

Wiegand and Klakow (2011a) compiled a list of prototypical opinion holders, i.e., common noun phrases such as *experts* and *analysts* that describe particular groups of people whose profession or occupation is to form or express opinions towards specific items. They assume that since those prototypical holders are common nouns, they should occur sufficiently often in a large text corpus.

Kim and Hovy (2006) used FrameNet data and semantic role labeling to extract opinion holders and scopes. They used a rich feature set of (1) keywords expressing opinion, (2) phrase types, (3) parse tree path, (4) position of the phrase (i.e. before or after the keyword), (5) the voice of the sentence (i.e. active vs. passive), and (6) the frame name. They obtained an F_1 score of 0.398.

Lu (2010) used dependency parsing to extract opinion holders in Chinese newswire texts. They relied on a number of heuristic rules such as: (1) in the case of reported speech, the subject of the reporting verb is the holder, (2) if the text is not reported speech, the text author is then the opinion holder; and (3) for news headlines if no reporting verbs are found, then the noun phrase before the colon is the holder, among other rules. Their approach yields an F_1 score of 0.784. Likewise, Rosá et al. (2010) proposed a rule-based system to extract opinion holders from Spanish newswire texts and achieved an F_1 score of 0.850.

Wiegand and Klakow (2011b) combined information about subjective expressions with Levin's verb classes to train a classifier to extract noun phrases in unambiguous agentive positions as the opinion holders. Their classifier gives a F_1 score of 0.650. They proposed that they can improve their classifier by restricting holder candidates to persons.

Compared to the aforementioned studies my approach to uncertainty attribution does not rely on prototypical holders or rules. I incorporate, instead, many features, including information about the predicted uncertainty cues, so that my machine learning model for uncertainty attribution can identify and extract new holder patterns beyond the ones in the training corpus.

4.4 Conclusion

In this chapter, I presented the second machine learning model in my comprehensive automatic uncertainty analyzer for Arabic tweets, namely uncertainty attribution. I used similar features to the ones used for uncertainty detection in Chapter 3. Yet, I added more new features. I used the same token sequence labeling methods that I used for uncertainty detection.

CHAPTER 5

UNCERTAINTY SCOPE EXTRACTION

5.1 Introduction

With two machine learning models for uncertainty detection and attribution in Chapters 3 and 4, respectively, the only remaining part of my comprehensive uncertainty analyzer for Arabic tweets is a machine learning model to identify and extract uncertainty scopes. Scopes are the linguistic constituents that encode the propositions modified by UCs. It is crucial for several NLP applications not only to identify and extract uncertainty cues and holders, but also to know what the uncertainty is about.

The remainder of this chapter is organized as follows: Section 5.2 describes my scope extraction approach: task description, classification features, experimental setup, results, discussion, and error analysis; and Section 5.3 compares my approach to others'.

5.2 Approach

5.2.1 Task Description

Following my unified framework, I cast uncertainty scope extraction as a token sequence labeling problem, in which the classifier predicts each token as the beginning of a scope

(B-S), inside a scope (I-S), or outside any scopes (O-S). Table 5.1 shows a corpus excerpt tagged with the BIO-S scheme.

Arabic	**Trans.**	**Gloss**	**English**	**BIO**
أنا	*>nA*	I	I	O-S
أرى	***>rY***	**see.1.sg.imprf**	**see**	**O-S**
أن	***>n***	**that**	**that**	**O-S**
الثورة	*Alvwrp*	the-revolution.sg.fm	the revolution	B-S
لن	*ln*	not	will not	I-S
تنتصر	*tntSr*	win.3.sg.fm.imprf	win	I-S
مادمنا	*mAdmnA*	as.long.as-we	as long as we	I-S
في	*fy*	in	are in	I-S
خلاف	*xlAf*	dispute	an ongoing	I-S
مستمر	*mstmr*	ongoing	dispute	I-S

Table 5.1: Uncertainty scopes represented in the BIO scheme and formatted according to the YamCha requirements

5.2.2 Classification Features

I add the Uncertainty Holder Features (UHFs) from Table 5.2 to all the previous features used for uncertainty detection and attribution. That is, for uncertainty scope extraction the feature set incorporates predictions from both uncertainty detection and attribution in addition to contextual, dialectal, semantic, lexicon, syntactic, pragmatic, and Twitter features.

No	Feature	Description
Uncertainty Holder Features (UHFs)		
1	holder	the text segment representing the holder of each identified cue
2	holder-location	whether each token comes before or after the identified holder in each tweet
3	holder-distance	the distance between each token and the identified holder in each tweet

Table 5.2: One more classification feature category for uncertainty scope extraction

5.2.3 Experimental Setup and Results

Similar to uncertainty detection and attribution, I use 10 fold cross validation to find the optimal (1) feature category combination using a greedy selection algorithm, (2) linear context width, (3) parsing direction, (4) multiclass method, and (5) polynomial degree.

Similar to uncertainty attribution, I use a simple bag-of-words model, based on token sequences around each given token, as a baseline model, which is again the same as the lexical contextual feature number 1 in Table 3.2. The baseline model performs worse for uncertainty scope extraction than it does for uncertainty attribution. This is expected given that linguistic structures encoding scopes are typically longer and more complex than those encoding holders.

Similar to both uncertainty detection and attribution, one-against-the-rest multiclass classification in a forward parsing direction (i.e. left to right) is the best configuration with the default YamCha polynomial kernel degree of 2. Although I mentioned earlier

No	Feature Category Combinations	Accuracy	Precision	Recall	F_1
0	Baseline	0.405	0.359	0.381	0.369
1	CFs	0.584	0.531	0.492	0.511
2	CFs+SynFs	0.618	0.574	0.528	0.550
3	CFs+SynFs+UCFs	0.640	0.610	0.584	0.597
4	CFs+SynFs+UCFs+UHFs	0.683	0.655	0.642	0.648
5	CFs+SynFs+UCFs+UHFs+SemFs	0.699	0.669	0.650	0.659
6	CFs+SynFs+UCFs+UHFs+SemFs+TFs*	0.702	0.678	0.653	0.665
7	CFs+SynFs+UCFs+UHFs+SemFs+TFs+PFs	0.705	0.677	0.661	0.669

Table 5.3: Results for uncertainty scope extraction with the best feature category combination marked with an asterisk

that in Arabic uncertainty scopes can precede or follow their cues, in my corpus, about 93.4% of the scopes follow their cues. As a result, the optimal right linear context width for uncertainty scope extraction is as large as +11; whereas the optimal left linear context width is only -3.

CFs, SynFs, and UCFs make it right away to the third round of the greedy feature selection algorithm similar to uncertainty attribution. UHFs, however, win the fourth round of the algorithm, mainly because UHFs combined with the UCFs eliminate even more tokens from being considered for scopes. The efficiency of the UCFs and UHFs is supported by the typical short length of tweets: once some tokens are labeled as encoding cues and others as encoding holders, a few tokens remain to be considered for scopes. This highlights one more time the advantage of the unified framework that I propose in this research project to identify and extract uncertainty cues, holders, and scopes in one fell-swoop.

SemFs and TFs introduce small improvements, with paired *t*-test *p*-values of 0.0342, and 0.0629, respectively. Significant performance improvements stop at the sixth round of the feature selection greedy algorithm with PFs giving insignificant improvement (*p*-value = 0.3079).

5.2.4 Error Analysis

Arranged based on their frequency, scope extraction errors include (1) scopes starting at the enclitic pronouns attached to their UCs, (2) syntactically complex scopes, typically comprising subordinate clauses, (3) scopes outside the sentence boundaries of their UCs, and (4) scopes outside the tweets of their UCs.

As I mentioned earlier, given that Arabic is an agglutinative language, the first parts of scopes can sometimes be enclitic object pronouns attached to the UCs as in example 1. Typically, the tokenizer should split those object pronouns. Yet, because the tokenizer I use, MADAMIRA v1.0 (Pasha et al., 2014), does not fully support Arabic dialects, many object pronouns go untokenized.

1.
    - العسكر فاكرينا هنخاف منهم.
    - *AlEskr **fAkry**nA hnxAf mnhm*.
    - the-army **thinks.3.pl.imprf-**us will-fear.3.pl.imprf from-them.
    - The army leaders **think** we are afraid of them.

Although tweets are typically short and syntactically simpler compared to texts from other linguistic genres, some tweets can include complex sentences as in example 2 that comprises two subordinate clauses, both of which are scopes for the UC أعتقد *>Etqd* (gloss: think.1.sg.imprf; English: I think).

2. - أعتقد أنه لن يتم التوصل لاتفاق بين قوى المعارضة حتى لو مر ١٠٠ عام حتى لو مات كل المصريين.
   - ***>Etqd >nh*** *ln ytm AltwSl lAtfAq byn qwY AlmEArDp HtY lw mr 100 EAm HtY lw mAt kl AlmSryyn.*
   - **think.1.sg.imprf that** not reach.3.sg.imprf.passive to-agreement among power the-opposition even if passed.3.sg.msc.prf 100 year even if died.3.sg.msc.prf all the-Egyptians.
   - I **think that** the opposition will not get to an agreement, even if they spend a 100 years trying, even if all Egyptians die.

Due to stylistic variations, scopes are not always in the same sentence of their UCs as in example 3. Furthermore, scopes are not always in the same tweet as their UCs. Tweets are like ongoing conversations among the users in which uncertainty information can be scattered across several tweets. However, inter-tweets scopes are only 500 cases in my corpus.

3. - يعني الإخوان هيسكتوا على حرق مقراتهم؟ … مظنش.
   - *yEny Al<xwAn hysktwA ElY Hrq mqrAthm? …* ***mZn$***.
   - meaning the-Brotherhood will-remain.silent.3.pl.imprf on burning headquarters -their … **not-think.1.sg.imprf-not**.
   - Is it that: the Brotherhood will not react against burning its headquarters? … I **do not think** (so).

5.3 Related Work

A number of scope extraction systems for English rely on manually-compiled lexico-syntactic rules, e.g., Kilicoglu and Bergler (2008, 2010); Ørelid et al. (2010); Rei and Briscoe (2010). Example rules are like: (1) the scope of a modal verb cue (e.g. may) is the verb phrase to which it is attached; and (2) the scope of a verb cue (e.g. seems) followed by an infinitival clause extends to the whole sentence. The F_1 scores of such systems range from 0.552 to 0.661 with all systems being trained and tested on English biomedical texts.

Manual creation of a comprehensive set of scope extraction rules is a laborious and time-consuming process. As a result, Apostolova et al. (2011) proposed deriving such rules automatically from a corpus annotated with speculation cues and their scopes, i.e., the BioScope corpus (Szarvas et al., 2008). Their approach achieves significantly higher F_1 scores of 0.756 for clinical papers, 0.789 for full papers, and 0.739 for abstracts.

Morante et al. (2010) dispensed with rule-based systems and used machine learning techniques to extract scopes from the BioScope Corpus, casting scope extraction as a token sequence labeling problem. Their classifier is trained to label each token as inside or outside a scope. They obtained an F_1 score of 0.809 with features such as word form, POS tags, chunks, types of chunks, and named entities. They used a post-processing algorithm to examine predicted discontinuous blocks of scopes and decide whether they should be combined or not. Similarly, Zhou et al. (2010) defined scope extraction as a token sequence labeling task and used CRFs for that purpose. They trained their classifier to decide for each given token whether it starts a scope, ends a scope, or none of these. Their feature set includes word, stem, chunk, and uncertainty cue features. They reported an F_1 score of 0.442. Both Morante et al. (2010) and Zhou et al. (2010) used post processing algorithms to find discontinuous scope chunks and join them together. According to the annotation guidelines of the BioScope Corpus they used, scopes are only continuous sequences of tokens.

5.4 Conclusion

In this chapter, I presented my machine learning model for uncertainty scope extraction which is the third and last model in my comprehensive uncertainty automatic analyzer. I used a rich feature set and obtained an F_1 score that is orthogonal with the results achieved in the NLP literature of scope extraction.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this research project, I presented a comprehensive automatic system for uncertainty analysis that is trained, tested, and evaluated for the Arabic language as used in the linguistic genre of tweets. The system comprises three machine learning models to identify and extract uncertainty cues, holders, and scopes in a pipeline fashion starting with cues and ending with scopes. My machine learning models yield F_1 scores of 0.839, 0.773, and 0.665, for uncertainty detection, uncertainty attribution, and uncertainty scope extraction, respectively. This makes the average F_1 score of the system 0.759.

The first contribution of my research project is that I work on an understudied uncertainty task, i.e., attribution, an understudied language, i.e., Arabic, and an understudied linguistic genre, i.e., tweets. Hence, I gain numerous insights. In the literature of Arabic NLP, the properties of Arabic as an agglutinative morphologically-rich language with a flexible word order are repeatedly claimed as challenges, not only for automatic uncertainty analysis as I do here in Sections 2.2.2.2 and 2.2.2.3, but also for other NLP tasks, including statistical machine translation (El-Kholy and Habash, 2012), parsing (Dehdari et al., 2011), and sentiment analysis (Abdul-Mageed et al., 2014), among many others. My empirical results, however, show that although such challenges do indeed exist and contribute to the errors of my machine learning models, they are not too pervasive to hinder Arabic NLP research or to yield poorly-performing machine learning models. For instance, in Section 2.2.2.2 I have mentioned that the rich morphology of Arabic packages information about uncertainty cues and holders in the same token as in معتقدش *mEtqd$* (gloss: not-think.1.sg.imprf-not; English: I do not think). Yet, the results show that only 7,992 out of 17,317 holders (i.e. 46.2%) are encoded in the morphological inflections of their UCs. In Section 2.2.2.3, I have also mentioned that agglutination can lead scopes to start at the enclitic object pronouns attached to their UCs. This is found to be true for only 4,232 scopes out of 16,817 (i.e. 25.2%). In the same Section 2.2.2.3, I have mentioned that Arabic long dependencies and flexible word order challenge both uncertainty detection and scope extraction. However, the results show that (1) 3,697 UCs are base phrase unigrams; (2) 10,544 out of the 13,620 multiword UCs are continuous multiword expressions; and (3) 15,708 out of 16,817 scopes immediately follow their cues. Therefore, my results show that the challenges frequently ascribed to Arabic do not represent the majority of the instances at least in my corpus. One reason for that might be the type of the linguistic genre I am working on, i.e., the genre of tweets. According to Badawi (2012), the more informal the linguistic genre is, the more simplistic semantics and syntax are used. That is why, according to Badawi (2012), we do not find many long dependencies or variable word orders in informal linguistic genres, such as tweets in my case. This raises an interesting future research question regarding the applicability of my machine learning models to more elaborate linguistic genres such as literary texts, scientific articles, and newswire texts, among many others.

The second contribution of my research project is that I use a unified framework based on sequence labeling methods to process the three tasks of uncertainty detection, attribution, and scope extraction. The unified framework enables using the predictions of one task to inform the other, and hence it boosts performance for some hard tasks such as scope extraction. Many researchers have worked simultaneously on both uncertainty detection and scope extraction. Yet, they have not used a unified framework for both tasks. Some researchers cast uncertainty detection as a token sequence labeling problem, and then use hand-crafted rules to extract scopes Ørelid et al. (2010); Yang et al. (2012); Apostolova et al. (2011); Velldal et al. (2010). Others define both uncertainty detection and scope extraction as token sequence labeling problems; yet use separate feature sets for each task or do not use the output of one task to inform the other Zhao et al. (2010). The importance of this unified framework is that it can be applied to other NLP tasks such as opinion mining and negation processing, for which researchers have also been interested to identify cues, holders, and scopes.

Table 6.1 shows examples of raw tweets and how predictions are added up as my pipeline proceeds from uncertainty detection to attribution and then to scope extraction. In the final output, tokens, that have been identified as uncertainty cues and as encoding the uncertainty holders in their morphology and/or semantics, are labelled as B-CH or I-CH, for Beginning of a Cue/Holder and Inside of a Cue/Holder. Furthermore, all tokens that have been labelled as O-Cue, O-Holder, and O-Scope are eventually given the label O-UC for O-Uncertainty to indicate that they do not encode any uncertainty information.

There are a few directions for future work that emerge from my research project. First, I have not investigated the performance of the features within each feature category. Second, I would like to test my machine learning models on different linguistic genres. Third, I used SVMs like many previous studies on uncertainty automatic analysis. Yet, it might be a good idea to compare SVMs with Conditional Random Fields (CRFs). Only Prabhakaran (2010) did that and found out that CRFs marginally improve the accuracy of the predictions, but substantially improve speed. Finally, I have built three different machine learning models for uncertainty detection, attribution, and scope extraction, and then arranged them in a pipeline fashion. An interesting future research question is whether the three models can be combined into one model.

Input				Tasks			Output
Arabic	Trans.	Gloss	English	Cues	Holders	Scopes	
اعرف	***AErf***	**know.1.sg.imprf**	**I know**	**B-C**	**B-H**	O-S	**B-CH**
ان	***An***	**that**	**that**	**I-C**	**I-H**	O-S	**I-CH**
مصر	*mSr*	Egypt	Egypt	O-C	O-H	B-S	B-S
في	*fy*	in	is in	O-C	O-H	I-S	I-S
مصيبة	*mSybp*	trouble	trouble	O-C	O-H	I-S	I-S
النظام	*AlnZAm*	the-regime	The oppressive	O-C	B-H	O-S	B-H
القمعي	*AlqmEy*	the-oppressive	regime	O-C	I-H	O-S	I-H
يظن	***yZn***	**thinks.3.sg.msc.imprf**	**thinks**	**B-C**	O-H	O-S	**B-C**
أن	***An***	**that**	**that**	**I-C**	O-H	O-S	**I-C**
الشعب	*Al$Eb*	the-people	the people	O-C	O-H	B-S	B-S
سينسى	*synsY*	will-forget.3.sg.msc.imprf	will forget	O-C	O-H	I-S	I-S
مش	*m$*	not	I do not	O-C	O-H	O-S	O-UC
عارفة	*EArfp*	know.1.sg.fm.imprf	know	O-C	O-H	O-S	O-UC
هيحصل	*hyHSl*	will-happen.3.sg.msc.imprf	what will	O-C	O-H	O-S	O-UC
إيه	<*yh*	what	happen	O-C	O-H	O-S	O-UC
متهيالي	***mthyAly***	**think.1.sg.imprf**	**I think**	**B-C**	**B-H**	O-S	**B-CH**
المقاطعة	*AlmqATEp*	the-boycott	the boycott	O-C	O-H	B-S	B-S
هي	*hy*	is	is	O-C	O-H	I-S	I-S
الحل	*AlHl*	the-solution	the solution	O-C	O-H	I-S	I-S

Table 6.1: Example output of my pipeline for uncertainty automatic analysis

REFERENCES

Abdul-Mageed, M., Diab, M., and Kübler., S. (2014). SAMAR: Subjectivity and Sentiment Analysis for Arabic Social Media. *Computer Speech and Language*, 28(1):20–37.

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford University Press, UK.

Al-Sabbagh, R., Girju, R., and Diesner, J. (2013). Using the Semantic-Syntactic Interface for Reliable Arabic Modality Annotation. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 410–418, Nagoya, Japan.

Al-Sabbagh, R., Girju, R., and Diesner, J. (2014a). *3arif*: A Corpus of Modern Standard and Egyptian Arabic Tweets Annotated for Epistemic Modality Using Interactive Crowdsourcing. In *Proceedings of the 25th Conference on Computational Linguistics (COLING 2014)*, pages 1521–1532, Dublin, Ireland.

Al-Sabbagh, R., Girju, R., and Diesner, J. (2014b). Unsupervised Construction of a Lexicon and a Repository of Variation Patterns for Arabic Modal Multiword Expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE) at EACL*, pages 114–123, Göthenburg, Sweden.

Alkuhlani, S. and Habash, N. (2011). A Corpus for Modeling Morpho-Syntactic Agreement in Arabic: Gender, Number, and Rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 357–362, Portland, Oregon.

Apostolova, E., Tomuro, N., and Demner-Fushman, D. (2011). Automatic Extraction of

Lexico-Syntactic Patterns for Detection of Negation and Speculation Scopes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 283–287, Portland, Oregon.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Avila, L. B. and Mello, H. (2013). Challenges in Modality Annotation in a Brazilian Portuguese Spontaneous Speech Corpus. In *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 1–6, Potsdam, Germany.

Azari, D., Dumais, S., Horvitz, E., and Brill, E. (2003). Web-Based Question Answering: A Decision Making Perspective. In *Proceedings of the Conference on Uncertainty and Artificial Intelligence*, pages 11–19, Acapulco, Mexico.

Badawi, E. (2012). مستويات العربية المعاصرة في مصر. دار السلام للطباعة والنشر والتوزيع والترجمة, القاهرة مصر.

Badawi, E., Carter, M., and Gully, A. (2004). *Modern Written Arabic: A Comprehensive Grammar.* MPG Books Ltd, UK.

Baker, K., Bloodgood, M., Dorr, B. J., Callison-Burch, C., Filardo, N. W., Piatko, C., Levin, L., and Miller, S. (2012). Modality and Negation in SIMT. *Computational Linguistics*, 38(2):411–438.

Brustad, K. E. (2000). *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian and Kuwait Dialects.* Georgetown University Press, Washington DC, USA.

Castillo, C., Mendoza, M., and Poblete, B. (2011). Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684, Heydrabad, India.

Cui, Y. and Chi, T. (2013). Annotating Modal Expressions in the Chinese Treebank. In *Proceedings of the IWC 2013 Workshop on Annotation of Modal Meaning in Natural Language (WAMM)*, pages 24–32, Potsdam, Germany.

de Marneffe, M.-C., Grimm, S., and Potts, C. (2009). Not a Simple Yes or No: Uncertainty in Indirect Answers. In *Proceedings of SIGDIAL 2009: the 10^{th} Annual Meeting of the Special Interest Group in Discourse and Dialogue*, pages 136–143, Queen Mary University of London.

de Marneffe, M.-C., Manning, C. D., and Potts., C. (2012). Did it Happen? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics*, 38:301–333.

Dehdari, J., Tounsi, L., and van Genabith, J. (2011). Morphological Features for Parsing Morphologically-rich Languages: A Case of Arabic. In *Proceedings of the 2^{nd} Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2011)*, pages 12–21, Dublin, Ireland.

Diab, M., Levin, L., Mitamura, T., Rambow, O., Prabhakaran, V., and Guo, W. (2009). Committed Belief Annotation and Tagging. In *Proceedings of the 3^{rd} Linguistic Annotation Workshop, ACL-IJCNLP'09*, pages 68–73, Suntec, Singapore.

Diab, M. T. (2009). Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. In *Proceedings of the 2^{nd} International Conference on Arabic Language Resources and Tools*, pages 285–288, Cairo, Egypt.

Díaz, N. P. C. (2013). Detecting Negated and Uncertain Information in Biomedical and Review Texts. In *Proceedings of the Student Research Workshop Associated with RANLP 2013*, pages 45–50, Hissar, Bulgaria.

El-Kholy, A. and Habash, N. (2012). Rich Morphology Generation Using Statistical Machine Translation. In *Proceedings of the 7^{th} International Natural Language Generation Conference*, pages 90–94, Utica.

Elfardy, H., Al-Badrashiny, M., and Diab, M. (2014). AIDA: Identifying Code Switching in Informal Arabic Text. In *Proceedings of the 1st Workshop on Computational Approaches to Code Switching*, pages 94–101, Doha, Qatar.

Elghamry, K., Al-Sabbagh, R., and ElZeiny., N. (2008). Cue-Based Bootstrapping of Arabic Semantic Features. In *ADT 2008: 9es Journes Internationales d'Analyse Statistique des Donnes Textuelles*, pages 85–95, Lyon, France.

Farkas, R., Vincze, V., Möra, G., Csirik, J., and Szarvas, G. (2010). The CoNLL-2010 Shared Task: Hedges and their Scope in Natural Language Text. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Uppsala, Sweden.

Goujon, B. (2009). Uncertainty Detection for Information Extraction. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, pages 118–122, Borovets, Bulgaria.

Habash, N. and Roth, R. (2011). Using Deep Morphology to Improve Automatic Error Detection in Arabic Handwriting Recognition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 875–884, Portland, Oregon.

Hendrickx, I., Mendes, A., and Mencarelti, S. (2012). Modality in Text: A Proposal for Corpus Annotations. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 1805–1812, Istanbul, Turkey.

Kilicoglu, H. and Bergler, S. (2008). Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective. In *Proceedings of BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, pages 46–53, Columbus, Ohio, USA.

Kilicoglu, H. and Bergler, S. (2010). Approach to Detecting Hedges and their Scopes. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 70–77, Uppsala, Sweden.

Kim, S.-M. and Hovy, E. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia.

Krippendorff, K. (2004). Measuring the Reliability of Qualitative Text Analysis Data. *Annenbery School of Communication: Departmental Papers, University of Pennsylvania.*

Li, J., Zhou, G., Wang, H., and Zhu, Q. (2010). Learning the Scope of Negation via Shallow Semantic Parsing. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 671–679, Beijing, China.

Lu, B. (2010). Identifying Opinion Holders and Targets with Dependency Parser in Chinese News Texts. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 46–51, Los Angeles, California.

Maamouri, M., Bies, A., Krouna, S., Gaddeche, F., and Bouziri, B. (2009). Penn Arabic Treebank Guidelines. *Linguistic Data Consortium.*

Marton, Y., Habash, N., and Rambow, O. (2013). Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features. *Computational Linguistics*, 39(1):161–194.

Matsuyoshi, S., Eguchi, M., Sao, C., Murakami, K., Inui, K., and Matsumoto, Y. (2010). Annotating Event Mentions in Text with Modality Focus and Source Information. In *Proceedings of the 7th International Conference Language Resources and Evaluation (LREC'10)*, pages 1456–1463, Valletta.

Mitchell, T. F. and Al-Hassan, S. A. (1994). *Modality, Mood, and Aspect in Spoken Arabic with Special Reference to Egypt and the Levant.* Kegan Paul International, London and NY.

Morante, R., Liekens, A., and Daelemans, W. (2010). Learning the Scope of Negation in Biomedical Texts. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 715–724, Honolulu, USA.

Mowery, D., Velupillai, S., and Chapman, W. (2012). Medical Diagnosis Lost in Translation – Analysis of Uncertainty and Negation Expressions in English and Swedish Clinical Texts. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012)*, pages 56–64, Montreal, Canada.

Ørelid, L., Velldal, E., and Oepen, S. (2010). Syntactic Scope Resolution in Uncertainty Analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1379–1387, Beijing, China.

Palmer, F. R. (1986). *Mood and Modality.* Cambridge University Press, Cambridge, UK.

Pasha, A., Al-Badrashiny, M., Diab, M., ElKholy, A., Eskandar, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: a Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland.

Prabhakaran, V. (2010). Uncertainty Learning Using SVMs and CRFs. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 132—137, Uppsala, Sweden.

Qazvinian, V., Rosengren, E., Radev, D. R., and Mei, Q. (2011). Rumor has it: Identifying Misinformation in Microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589—1599, Edinburgh, Scotland, UK.

Rei, M. and Briscoe, T. (2010). Combining Manual Rules and Supervised Learning for Hedge Cue and Scope Detection. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 56–63, Uppsala, Sweden.

Rosá, A., Wonsever, D., and Minel, J.-L. (2010). Opinion Identification in Spanish Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 54–61, Los Angeles, California.

Rubin, V. L. (2007). Stating with Certainty or Stating with Doubt: Inter-Coder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In *Proceedings of NAACL HLT 2007, Companion Volume*, pages 141–144, Rochester, NY.

Rubinstein, A., Harner, H., Krawczyk, E., Simoson, D., Katz, G., and Portner, P. (2013). Toward Fine-Grained Annotation of Modality in Text. In *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meaning in Natural Language (WAMM)*, pages 36–46, Potsdam, Germany.

Ruppenhofer, J. and Rehbein, I. (2012). Yes we can!? Annotating the Sense of English Modal Verbs. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 1538–1545, Istanbul, Turkey.

Saurí, R. and Pustejovsky, J. (2009). FactBank: A Corpus Annotated with Event Factuality. *Language Resources and Evaluation*, 43:227–268.

Shaalan, K., Bakr, H. A., and Ziedan, I. (2009). A Hybrid Approach for Building Arabic Diacritizer. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 27–35, Athens, Greece.

Soni, S., Mitra, T., Gilbert, E., and Eisenstein, J. (2014). Modeling Factuality Judgments in Social Media Text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 415–420, Baltimore, Maryland, USA.

Szarvas, G. and Gurevych, I. (2013). Uncertainty Detection for Natural Language Watermarking. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 1188–1194, Nagoya, Japan.

Szarvas, G., Vincze, V., Farkas, R., and Csirik, J. (2008). The BioScope Corpus: Annotation for Negation, Uncertainty and their Scope in Biomedical Texts. In *Proceedings of BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio.

Szarvas, G., Vincze, V., Farkas, R., Möra, G., and Gurevych, I. (2012). Cross-Genre and Cross-Domain Detection of Semantic Uncertainty. *Computational Linguistics*, 38:335–367.

Szolovits, P. (1995). Uncertainty and Decisions in Medical Informatics. *Methods of Information in Medicine*, 34:111–121.

Täckström, O., Eriksson, G., Velupillai, S., Dalianis, H., Hassel, M., and Karlgren, J. (2010). Uncertainty Detection as Approximate Max-Margin Sequence Labelling. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 84–91, Uppsala, Sweden.

Tang, B., Wang, X., Wang, X., Yuan, B., and Fan, S. (2010). A Cascade Method for Detecting Hedges and their Scope in Natural Language Text. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 13–17, Uppsala, Sweden.

Tjong, E. and Sang, K. (2010). A Baseline Approach for Detecting Sentences Containing Uncertainty. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 148—150, Uppsala, Sweden.

Velldal, E., vrelid, L., and Oepen, S. (2010). Resolving Speculation: MaxEnt Cue Classification and Dependency-Based Scope Rules. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 48–55, Uppsala, Sweden.

Vincze, V. (2013). Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 383—391, Nagoya, Japan.

Vincze, V. (2014). Uncertainty Detection in Hungarian Texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014): Technical Papers*, pages 1844–1853, Dublin, Ireland.

Vincze, V., Szarvas, G., Móra, G., Ohta, T., and Farkas, R. (2011). Linguistic Scope-Based and Biological Event-Based Speculation and Negation Annotations in the BioScope and Genia Event Corpora. *Journal of Biomedical Semantics*, 2(5).

Vlachos, A. and Craven, M. (2010). Detecting Speculative Language Using Syntactic Dependencies and Logistic Regression. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 18–25, Uppsala, Sweden.

Wagner, C., Liao, V., Pirolli, P., Nelson, L., and Strohmaier, M. (2012). It's not in their Tweets: Modeling Topical Expertise of Twitter Users. In *Proceedings of 2012 ASE/IEEE International Conference on Social Computing*, pages 91–100, Washington DC, USA.

Wei, Z., Chen, J., Gao, W., Li, B., Zhou, L., He, Y., and Wong, K. (2013). An Empirical Study on Uncertainty Identification in Social Media Context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 58–62, Sofia, Bulrgaria.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2):163–210.

Wiegand, M. and Klakow, D. (2011a). Prototypical Opinion Holders: What We can Learn from Experts and Analysts. In *Proceedings of Recent Advances in Natural Language Processing (RANLP'11)*, pages 282–288, Hissar, Bulgaria.

Wiegand, M. and Klakow, D. (2011b). The Role of Predicates in Opinion Holder Extraction. In *Proceedings of the Workshop on Information Extraction and Knowledge Acquisition*, pages 13–20, Hissar, Bulgaria.

Yang, H., Roeck, A. D., Gervasi, V., Willis, A., and Nuseibeh, B. (2012). Speculative Requirements: Automatic Detection of Uncertainty in Natural Language Requirements. In *Proceedings of 20th IEEE International Conference on Requirements Engineering*, pages 111–20, Chicago, IL, USA.

Zhao, Q., Sun, C., Liu, B., and Cheng, Y. (2010). Hedges and their Scope Using CRFs. In

Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task, pages 100–105, Uppsala, Sweden.

Zhou, H., Li, X., Huang, D., Li, Z., and Yang, Y. (2010). Exploiting Multi-Features to Detect Hedges and Their Scope in Biomedical Texts. In *Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task*, pages 56–63, Uppsala, Sweden.