Annotating Attribution Relations in Arabic

Amal Alsaif, Tasniem Alyahya, Madawi Alotaibi, Huda almuzaini, Abeer Algahtani

Al-Imam Mohammad Ibn Saud Islamic University
College of Computer Sciences and Information
{asmalsaif, tnalyahya, msalotaibi, hamozeani}@imamu.edu.sa;aakalqahtani@sm.imamu.edu.sa}

Abstract

We present a first empirical effort in annotating attribution in Modern Standard Arabic (MSA). Identifying attributed arguments to the source is applied successfully in diverse systems such as authorship identification, information retrieval, and opinion mining. Current studies focus on using lexical terms in long texts to verify, for example, the author identity. While attribution identification in short texts is still unexplored completely due to the lack of resources such as annotated corpora and tools especially in Arabic on one hand, and the limited coverage of different attribution usages in Arabic literature, on other hand. The paper presents our guidelines for annotating attribution elements: cue, source, and the content with required syntactical and semantic features in Arabic news (Arabic TreeBank - ATB) insight of earlier studies for other languages with all required adaptation. We also develop a new annotation tool for attribution in Arabic to ensure that all instances of attribution are reliably annotated. The results of a pilot annotation are discussed in addition to the inter-annotators agreement studies towards creating the first gold standard attribution corpus for Arabic.

Keywords: attribution, annotation tool, NLP, Arabic discourse, annotation guidelines, ATB, inter-annotator agreement.

1. Introduction

Textual information is one of the huge significant data available in the World Wide Web (WWW), which rapidly spread globally. This information is versatile and reflects different opinions as well as behaviours. For example, most of news is reporting people's speech and opinions about particular events, with additional explanation and analysis by the writer to tend people into desired understanding and background. Referring attributed arguments to the source and distinguishing different opinions than the author/writer opinion are not straightforward processes (Pareti 2012). Attribution could be direct speech (quotations) with no influence by the author on the content, or indirect speech when the abstract object is presented in a different way than the exact speech or it is not clear who state the argument. Usually quotation tools are used in direct speech such as punctuations (: and "), speech acts such *say/qAl/قال* or *comment/Akbar/أخبر*/and some particles such as that/An/أن, as in Example 1. On other hand, indirect speech may miss one or more of these tools which leads to ambiguity on determining the exact source and content boundaries, Examples 2 and 3. This makes identifying quotations and opinions of other people require advanced analysis and tools in addition to the basic lexical-syntactic analysis in the literature (Pareti 2012).

(1)
قال مدرب مانشستريونايتد مورينيو [بخصوص المباراة]:"*كانت حصة تدريبية جيدة جدا* ".
Manchester United coach Mourinho **said** [about the match]: "*it was a very good training share"*

(2)
علق مصدر مسؤول في الحكومة {رفض ذكر اسمه} [على تصريح الرئيس المتضمن رفع اسعار البترول] *أن هذا القرار من شأنه زعزعت الدخل العام للاسر*.
An official source in the government{refused to mention his name} **commented** [on the president's statement, which included raising oil prices], *that this decision would affect public income for families*

(3)
أكد السويسري كريستيان جروس {الذي يعتبر مدرب الفريق الأول لكرة القدم}، [عند اجتماعه باللاعبين]**على** *أهمية المباراةاليوم*.
Swiss coach Christian Gross, {who is considered the coach of the first football team}, he **confirmed** [when meeting with the players**] on** *the importance of the game today*.

Attribution annotation has recently received significant attention in Natural Language Processing (NLP) due to its relevance in particular to information extraction; question answering, story generation, summarization, and opinion mining (Guzmán-Cabrera et al. 2009; Juola and Baayen 2005; Neumann and Schnurrenberger 2009; Wiebe 2002). Most of these studies dealt with attribution in English. However, to the best of our knowledge, there are no empirical studies on annotating attribution in Arabic to generate a gold standard corpus. We propose in this paper our approach of identifying attribution in News corpus of contemporary Arabic, the Arabic Treebank (Maamouri et al. 2004) of both direct and indirect quotations. The work is inspired by a large discourse annotation project, the PDTB (Prasad et al. 2006; Prasad et al. 2007) that annotate attribution that serve discourse relations in English. The attribution annotation was extended by Silvia and her colleagues (Pareti et al. 2013) in annotation attribution regardless the existing of discourse relations.

The rest of the paper is structured as follows: Section 2 overviews the popular similar attribution annotation efforts and related applications. Section 3 presents a brief review about attribution in Arabic literature. Our schema of annotating the main elements and semantic features of attribution is discussed in section 4. ESNAD annotation tool and the human annotation process are discussed in Section 5. Then, Section 6 presents the pilot annotation of 20 news articles from different resources with a discussion about the agreement studies between annotators. We conclude with observations and actions required for creating the first gold standard corpus of attribution in Arabic.

2. Related work

Over the literature, there have been several studies that addressed the attribution in computational linguistics, mostly for English. These studies varied on considering one or more aspects of a lexical, syntactical, and semantic structure of automatic quotations, with very limited effort on indirect quotations. For example, work performed by (Mamede and Chaleira 2004; Elson and McKeown 2009) on narrative text and (Pouliquen, Steinberger, and Best 2007; Sarmento, Nunes, and Oliveira 2009; Schneider et al. 2010) on news text are based on lexical terms and some syntactic rules to infer the author of the quoted text. Experimental evidence in these studies clearly indicates the unreliability on identifying all attribution instances. Later, the work by (Elson and McKeown 2010; Fernandes, Motta, and Milidiú 2011; O'Keefe et al. 2012) had better success due to the fact they were based on NLP approaches such as rule-based and statistical machine learning of syntactical structure feature. Other studies consider specific types of attribution such as opinions at sentence level in the Multi-Perspective Question Answering MPQA (Wiebe 2002). The set of features used in annotating the MPQA includes: on, inside (content), and outside. The source is not annotated independently; it is labelled as 'outside' together with everything in the sentence other than the cue and the content.

The limitation of used data limits feature extraction too. (Elson and McKeown 2010) used a corpus consisting of about 3,000 quotes, and manually identified candidate features of the speaker mainly their gender and attributes each quote to the most likely speaker. This proposed approach achieved an average accuracy of 83%. In (Fernandes, Motta, and Milidiú 2011), they used a corpus contains annotation of entities, co-references, quotes, associations between quotations and authors, and part-of-speech tagging features. The performance achieved for author attribution was 79,02%. While the dataset in (O'Keefe et al. 2012) used the same corpus of (Elson & McKeown, 2010) and adds about 5,000 attributions from the PDTB which is news from the Wall Street Journal (WSJ) and Sydney Morning Herald (SMHC) that has been annotated with over 3,500 direct quotations and their speakers, but the cue element is not annotated since the cue can be inferred implicitly. The accuracy of this system was 84,1% and 91,2% for WSJ and SMH respectively. As concluded by (Pareti 2016), there is a lack of a comprehensive theory of attribution and a large gold standard annotated corpus with all attribution elements and features, which clearly influence on a performance of machine learning systems to identify attribution elements: the quote, its source, the purpose of reporting, its cue and its truthful level.

A part from PARC3 corpus (Pareti 2016), attribution was not the core of most annotated corpora. It was integrated in a limited extend with other discourse phenomena such as factuality in FactBank and discourse relations in RST and the PDTB. In the FactBank (Saurí and Pustejovsky 2009), the attributed span itself is not marked, but its events are linked to their source by introducing predicates in order to derive their factuality. Consequently, the annotation schema resulted in a relatively high inter-annotation agreement %81. The work in (Carlson and Marcu 2001) on the RST discourse treebank and (Wolf and Gibson 2005) the GraphBank projects consider attribution as a discourse relation. However, the first annotates only intra–sentential attributions with an explicit source and a verb cue, the latter annotates attribution if no other discourse relation is present. Inter-annotation agreement was evaluated and reported in all three copra. FactBank have resulted in a relatively high inter-annotation agreement (kappa=0.8). The results of the inter-annotator agreement within the RST framework was tested by multiple judges during multiple phases of the development of the RST corpus. Kappa values of the RST and the GraphBank reflects considerably high levels of agreement, greater than 0.8.

The large existing resource for annotating attribution as inter-sentential discourse relation is in the PDTB project (Prasad et al. 2006; Prasad et al. 2007) with 10K annotation of attribution. Attribution is a relation between abstract object and the source entity that must relate to one of discourse relations. However, this approach leaves out several instances of attribution and therefore some related features, e.g. nested attribution with no annotation. They annotate features for each accepted attribution in their scheme such as: type, source, determinacy, and polarity; Where the type may indicate one of the four distinct sub-types: assertion proposition, belief proposition, facts and, eventualities. An analysis of inter-annotator agreement was conducted on the PDTB corpus. The high inter-annotator agreement achieved indicates that discourse connectives and their arguments expose a well-defined level of discourse structure that can be reliably annotated.

The PDTB paradigm was applied similarly to other languages such as Chinese (Zhao and Zobel 2005; Huang and Chen 2011; Zhou and Xue 2012; Zhou et al. 2014; Zhou and Xue 2015), Arabic (Al-Saif and Markert 2010), Hindi (Kolachina et al. 2012), Czech (Mírovský, Jínová, and Poláková 2014), and Turkish (Zeyrek and Kurfalı 2017). Attribution in all these studies is not embraced currently in the annotation scheme and left to be extended in the future. In (Li et al. 2014) a Connective-driven Dependency Tree (CDT) structure as a representation scheme for Chinese discourse structure is proposed. CDT takes advantage of both RST and PDTB, and well adapts to the special characteristics of Chinese discourse relation including attribution. Later in (Kong and Zhou 2017), a CDT-styled end-to-end Chinese discourse parser was developed.

Further extension of the PDTB to annotate attribution was proposed by Pareti and her colleagues to annotate direct and indirect attribution in a comprehensive coverage in the Italian Attribution Corpus (ItAC) (Pareti and Prodanof 2010) and the Penn Attribution Relation Corpus (PARC3) (Pareti 2016). The ItAC is a small-scale Italian pilot corpus (461 instances in 50 articles) annotating key features of attribution: the source, cue, content, supplement, and additional features. Still, the schema needs to be tested for inter-annotator agreement. PARC 3.0 is the first large English corpus fully annotated with 19,712 attribution relations. It is initially grounded on the annotation in the PDTB, in addition to annotating new instances not related to discourse relations and annotate nested attributions. The attribution components include: source, cue, content, and supplement with a set of features included in the PDTB: attribution type, source type, factuality, and scopal polarity. While the annotation schema proposed in ItAC has not yet

been validated by inter-annotator agreement, the inter-annotator agreement results for the annotation of the spans in PARC 3.0 corresponding to source, cue, content and supplement are reported to be 100%, 91%, 94%, and 46% respectively. The ItAC and PARC3 are the only corpora have allowed the identification of several attribution structures not addressed by former studies.

Among a few studies of authorship identification in Arabic (Ouamour and Sayoud 2013; Rabab'ah et al. 2016) a recent study that addressed Arabic authorship identification on short text was conducted by (Rabab'ah et al. 2016) on 38,386 tweets for 12 users. Using SVM classifier with lexical and syntactic features, their system achieved accuracy of 68.67% on assigning each tweet with corresponding author. This study focused on the quote (tweet) and the source (the user) only with no use of other attribution elements and features, or nested attribution.

3. Attribution in Arabic

Arabic is categorised into either classic Arabic (CA) or contemporary Modern standard Arabic language (MSA). Both are sharing main characteristics of Arabic morphology, syntax and semantic. The differences lay on the usage and the level of construction with new vocabularies throw generations. MSA is used nowadays in books, education, news, but not always used in spoken language due to the effects of dialects of different regions (Habash 2010). Arabic NLP studies use mostly MSA, especially in information analysis and retrieval. Arabic processing requires deep multi-layer text processing in tokenization, stemming, morphological, syntactical analysis, and discourse (Farghaly and Shaalan 2009). In Example 4, the verbal sentence 'then they will read it' is represented in Arabic as one white spaced word (فسيقرأونها/ fsyqrAwnha/then they will read it) with many clitics: one proclitic, one prefix, one postfix and one enclitic are all attached to the stem يقرأ/yqrA/read. Moreover, most written text in MSA lacks using punctuations and diacritics (long vowels) in standardized manner, leading to high ambiguity in any automatic text processing.

(4) Gloss and syntactic analysis of a token
فسيقرأونها / then they will read it:

ف /then (connective)
+س / will (tense)
+ يقرأ /read (present verb)
+ ون/they (subject)
+ ها /it (object)

Traditional literature in semantic science/علم المعاني discusses speech reporting (أسناد /نقل الكلام/رواية) as one of the writing styles and study its basic structure using lexical, syntactic, or pragmatics features not connected to any discourse theories (Ali 2004). Similar to other languages, reporting others speech may use direct (exact words) or indirect quoting (not exact words but keeping the semantic). The content is usually enclosed in quotation or (:) marks to be direct, or precede by adverb (أن/An/that) to be indirect quoting. In fact, indirect reporting has many forms: changing lexical words and syntax of the original text, may use (أن/An/that) or not, may have two explicit verbs (speech act and a verb in the verbal phase in the content), or one verb serving the two verbs purpose with implicit declaration speech act. The writer shows his style when reporting other speech using indirect quotation, he may emphasise on some points or to tend the reported argument to be evidence to specific claims, which sometimes differ that the source intention and so making rumours. For example, He said "I'm going early" is a direct reporting, whereas He said (that) he was going early is indirect speech reporting. Modern Arabic studies barrow some theories from other languages that focus on the usage and semantics of reporting speech, and how relate to the source and its writing style. For example, the famous debatable theory "the Speech Acts theory/نظرية أفعال الكلام" by Searle (Searle 1976) that was adapted from Austin theory (Austin 1975). They claimed that speech acts are used to express the purpose of the quoting by the author. For example, the semantic of say/قال is different than assure/شدد even though the quote is identical. According to Searle, there are two types of Speech Acts: *constatives (declaratives) and performatives*. Constatives are used for propositions which can be stated to the truth value. For example, he said: *sky is blue/السماء زرقاء* :قال the proposition presents a status which can be assessed to the truth at that time with no action by the source.

Performative verbs express specific action belong to a purpose such as: Assertives (i.e. suggest/AqtrH /اقترح), Directives (i.e. ask/اسأل), Commissives (i.e. promise/ wEd /وعد) in Example 5, Expressives (i.e. forgive/AEtdr/ /اعتذر), and Declaratives (i.e. confirm/Akd/أكد). A paper by (Bouayad and Belkheer 2012) presents a collection of most frequent speech acts in Arabic. However, there is no classification of direct/indirect speech acts in use, and no clear distinguishing between implicit and explicit attribution, some books mix them with direct/indirect speech.

(5)

أعدك *أن أحضر باكرا .*

*I **promise** you that I will come early*

Although, these studies criticise that Speech Acts theory is incomplete and speech acts may lie under more than one class, there is no special Arabic theory would cover this lack to be used in a complete empirical study of attribution. Further, there is no guideline for annotating the quotations (direct and indirect) in textual corpus that could be used on machine learning modelling. This highlights the importance of analysis of a corpus-based study of attribution in Arabic, which might be used to prove and extract new classification of attribution types and analyse new indirect reporting styles in use.

4. Attribution schema of Arabic

This section describes our first attempt in annotating attribution relation in MSA. The ground base of the proposed annotation is the work of the PDTB (Prasad et al. 2006) and Silvia in (Pareti and Prodanof 2010; Pareti 2011; Pareti 2012) but tailored and extended to suit the MSA. We share with other studies the interest in identifying the basic elements of attribution: cue, the source, and the content. Authors while paraphrase the others speech may use temporal cues and locations to increase the truth of their reporting or describing the entity in more details, so called supplement information. Each basic elements may have

supplement. The author, also, may use implicit indirect speech and increase the ambiguity in determining the source of the claim, the author or others. The schema is broken down into four main classes with clear definition and examples of each: the cue, attribution general features, source, and content. Each class has further features to cover all semantic aspects.

Examples in this paper are presented according to the following convention: the cue is (bold-faced), the source (underlined), the content (italic), the cue supplement (enclosed in brackets), the source supplement (enclosed in braces), and the content supplement (enclosed in parentheses).

4.1 Cue

The cue can be defined as the lexical anchor that links the source with the content. The cue may occur in different syntactic forms: a reporting verb/speech act either *constatives* تصريحية or *performatives*انجازية (such as emphasis/Akd/أكد in Example 3), adverb (such as adding/ mDyfA/مضيفا in Example 6), an adjective (such as describing/واصفا), and a prepositional phrase (such according to/بحسب). The cue also can be omitted such as (Ahmad: I will go/أحمد: سوف أذهب) which is understood from the context as *said "…"*.

(6)
مضيفا *" هذا ما فكرنا به عندما كنا نخطط لاستعدادات الموسم الجديد "*.
He **added** *"That's what we thought of when we were planning for the new season."*

Cue status is a feature indicates whether the cue is explicitly occurred in the text (explicit) or omitted (implicit). *Explicit* cue is usually one of the declarative reporting acts (such as say/قال, mention/ذكر and declare/صرح), with direct speech when the content is introduced by punctuation marks (: or ""), or with indirect speech when it is not clear whether the content has exact words of actual speech. The particle (that /ان) could be used for indirect speech. On other hand, *implicit* cue is a feature when the cue is either omitted or being a *performative* speech act with indirect speech only such as (*he thanks his teacher/شكر معلمه, he suggests not to answer/ اقترح عدم الاجابة, he promised to his dad coming early/ وعد اباه العودة باكرا*). The reason behind counting these speech acts as implicit cue is that ability of converting them into explicit cues by adding one of reporting acts which are declaration acts (such as say/قال, mention/ذكر, declare/صرح) and the exact speech in the content but the writer preferred to make them implicit. For example, *he said thankfully to his teacher "…."/…. :قال شاكرا لمعلمه, he said I suggest to not answer/قال اقترح عدم الاجابة and he said to his dad I promise to come early/قال واعدا أباه انه سيعود باكرا*). Examples 1 has explicit cue with direct speech, Example 2 has explicit cue with indirect speech and while Example 3 has implicit cues with indirect speech.

Cue supplement is the text span that describe the status of the cue (when?where?how?) which is relevant to the interpretation of the cue such as adverb *laughing/ضاحكا*, temporal phrase, or place of the attribution cue. Some modifiers such as prepositions as *on/على , text on/ نص على* are tagged as part of the cue, not as supplement.

Cue negation determines if the cue is modified by a negation tools (e.g., did not/lm/ لم) or the cue itself indicates negation semantically (i.e *denied/rfD/رفض*).

Cue digression when the cue digresses a former quotation such as in Example 6 where the attribution could not be stand alone at the first place.

4.2 Source

The Source is the entity holding the content. As in the PDTB (Prasad et al. 2006; Prasad et al. 2007) , the source is annotated by marking a text span expressing the source and also its type. Syntactically the source will be the subject of the declaration verb either in explicit or implicit cue. The source type express diverse types of agents: (i) the writer of the text (Mustafa 2011) - the writer is reporting someone lese speech directly, (ii) any specific agent other than the writer either explicitly occurred in the text (EXP-AG) such as in Example 1, (iii) the source in not explicitly appear in the same sentence of the attribution but could be inferred from previous context, this tag as an implicit agent (IMP-AG) such as in Example 6, (iv) or the source is anonymous and the writer did not refer this speech to a specific agent (Miss) as in Example 7 . Our new feature here is **the source supplement** to tag any expression or relative clauses related to the source, sometimes the writer prefers to describe the agent in more details as in Examples 2 and 3.

(7)
يذكر *أن السعودية تواصل لليوم الثالث تسيير رحلاتها المتتابعة داخلياً ودولياً بكل انسيابية.*
It is **said** that *Saudi Arabia continues for the third day running its successive flights internally and internationally smoothly.*

4.3 Content

The attributed material is annotated as **a content feature,** by determining the text span boundaries that might range from only one word into multiple sentences. The content may cover the cue too if the cue plays as basic verb in the content such as in implicit indirect attribution (Khaled congratulated his brother on success/هنأ خالد أخيه على النجاح) **Content supplement:** any clauses the writer added to the content to present some background information about the entities or events in the content but is mostly not part of the reported speech. For example, the relative clause in Example 8 *those who use their mobiles while driving/ الذين يستخدمون الجوال أثناء القيادة* is part of the content and not supplement. **Content negation**: the content or attributed speech is negated when using either negation function words such as *did not/لم and will not/لن*as in (Ahmad said I will not go to the school/قال أحمد لن أذهب للمدرسة) and using a noun *none/عدم* in Example 9, or using negation verbs or nouns such as refusal/الرفض and denial/انكار in the content itself.

(8)
طالب السيالي[إدارة المرور] *بملاحقة مشاهير السناب شات الذين يستخدمون الجوال أثناء القيادة.*
Al-Sayali **requested** [the Traffic Department] *to track down the snap celebrities* who use mobile phones while driving.

(9)
ذكر الحمادي: *عدم تأهل نادي الهلال لبطولة أسيا.*

Al-Hammadi **said**: *Al-Hilal is not qualified for the Asian Championship.*

4.4 General features

Apart from the text span features (cue, source, content and their supplements), the attribution relations have further key semantic features that might be used in discourse processing and information/opinion extraction systems. The general features include: *attribution style, determinacy, and attribution purpose.* **Attribution style** distinguishes whether the content is quoted with exact words of the spoken (direct) as Example 1, or reports someone else speech without using that person's exact words (indirect) as in Example 2. Implicit cues often use indirect speech, while explicit cues could use either direct or indirect. **Determinacy feature**, was borrowed from the PDTB, it identifies the factuality of the attribution relation itself; not the content. The feature will be indeterminacy (Non-Factual) when the relation in the scope of hypothesis, negation, or future tense such as using (will/سوف or may-perhaps/ربما) as modifier to the cue, see Example 10. In hypothesis, the conditional function words such as *if* /لو is used to express the factuality of attribution, as in Example 11.

(10)

إن فتشنا في الملفات ربما **يقول** الجيل الجديد *أين الرقابة، وأين المسؤول، وأين الإعلام؟*

If we look at the files, the new generation might **say**, "*Where is censorship, where is the administrator, and where is the media?"*

(11)

لو **أخبرنا** *أنه سيعبر أجواءنا ،كنا سنحتفي به.*

If he **told** us that *he will pass our atmosphere*, we would celebrate him.

Attribution purpose, while the purpose of the speech reporting is transfer the news, facts, or stories, this feature signifies the nature of the relation between an agent and the cue, it describes the reason of using this particular cue in reporting the speech. Our annotation guidelines tried to base on well-established linguistic theories as possible. As mentioned earlier we use the classification of speech acts (cues) in the Speech Acts Theory (Searle 1976) and it is application on Arabic to determine the purpose of the reporting. Our taxonomy has flat distribution of five distinct purposes of attribution: (i) **Assertion** when commit to the truth of the proposition (e.g. said/قال , assert/أكد/جزم, mention/dkr/ذكر), (ii) **Directive** for requesting (ask/سأل, order/أمر or request/طلب) or questioning (question /استفهم), (iii) **Expression** purpose is to express a feeling or regretting (e.g. apologies /اعتذر, congratulate/هنأ), (iv) **Declarative** to declare changing on affairs (e.g. announced/ أعلن, informed/أبلغ, admitted/اعترف and stated/أفاد). (v) **Commissive** acts express any commitments **(**e.g. bet/راهن, promise/وعد, and oath/ أقسم). Assertion and Declarative speech acts often used to express explicit attribution direct/indirect. While other purposes (Directive, Expression, and Commissives) are commonly used for implicit indirect attribution.

5. Building the first attribution corpus for Arabic

5.1 Corpus

We planned to enhance the discourse layer in the Leeds Arabic Discourse TreeBank (LADTB) (Al-Saif and Markert 2010), the valuable discourse resource for Arabic, by annotating explicit, implicit, direct and indirect attribution in 530 news articles from Arabic treebank ATB-Part1 (Maamouri et al. 2004). The ATB has morphological and syntactical gold standard annotation, used in many studies in Arabic NLP community. Adding our annotation to this corpus will encourage further studies on computational linguistics.

5.2 Annotation tool

Annotation tools share the graphic-based visualization that inspire users to gather complex annotations in an easy and reliable technique. While there are some general purpose annotation tools such as the GATE tool (Ide and Suderman 2009), BRAT tool (Stenetorp et al. 2012), MMAX2 tool (Müller and Strube 2006), and WebAnn (Yimam et al. 2013), few of them support relational annotation and Arabic calligraphy. We therefore, decided to develop a new Java-based annotation tool for Arabic (ESNAD: Extracting Sentence Attribution in Arabic Discourse) with a user-friendly interface to ensure highly reliable annotation which could be used for similar languages such as Urdu, see Figure 1.

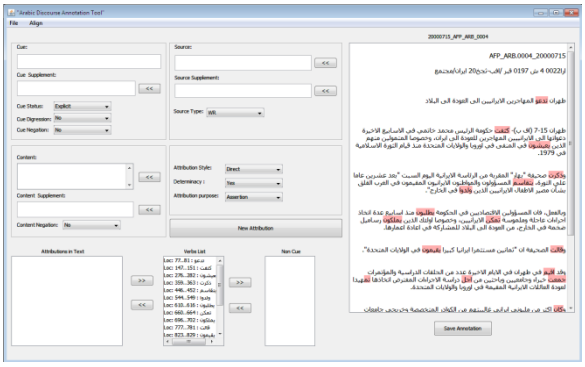

Figure 1: The main interface of ESNAD with initial highlighting of all verbs.

5.3 Human annotation process

The ESNAD tool will highlight all verbs (extracted from the ATB) and marked them as potential attribution cues. The annotator distinguishes between verbs presenting speech acts or not which are therefore not attribution. The annotator still has an ability to mark any cue such as adverbs or clauses from text on the right. The annotator should follow the annotation schema presented in Section 4 and annotate basic attribution elements and their features for direct/indirect quotation and explicit/implicit cue. We designed the tool to prevent any data entries by the user: so either marking desired text span on the text on the right side, or selecting a label from predefined labels as in our scheme. The tool saves all annotation into a text file to conduct the inter-annotator agreement and produce a gold standard annotated corpus by expert verification of disagreed cases. We conduct a pilot annotation study on 20

news articles from ATB-Part1 and other news websites to validate our guidelines. Annotators are native Arabs with a good linguistic background. The distribution of our annotation and observations are discussed in the next section.

6. Pilot annotation study

The annotation was conducted by 2 annotators (4 different Arabic speakers, 10 files each pair) after 2 rounds training on different 5 files using the tool and the schema. The inter-annotators agreement is measured using observed agreement, f-score and kappa to take into account agreement by chance (Siegel and Castellan 1988), as in Table 1. Among 734 potential instances, annotators agreed on 98% (161 as attribution cases and 534 as not attributed verbs). A high reliable annotation is recorded for the cue, its status (implicit/explicit) and attribution style (direct/indirect). A good agreement is recorded for attribution purpose and source type (~85%). The low kappa of the source type turns our attention on a high agreement by chance. This result is justified by being the default label of the source is the writer in the tool for each instance and the annotator should change it when appropriate, but he missed doing so for this feature. As a result, we will make this feature without default value in the tool.

	Accuracy	**F-score**	**Kappa**
Attribution Cue (734)	98%	0.96	0.95
Attribution Style (161)	94%	0.96	0.88
Cue Status (161)	94%	0.96	0.80
Attribution Purpose (161)	85%	0.79	0.76
Source Type (161)	83%	0.76	0.61

Table 1: inter-annotator agreement on label features

For text features we use the agreement measurement agr; it is introduced in (Wiebe, Wilson, and Cardie 2005)and used in many annotation studies. As in Equation 1, *agr* is an average of agr-annotror1 and agr-annotator2 when each one is calculated by dividing number of matching words of the two text annotations by the total number of words of that annotation.

*agr=1/i *∑((# of matching words)/(# of total words in ann(i)*

Table 2 shows highly reliable agr agreement of all text features with higher that 96%. Not surprisingly that supplement features of (cue, source and content) have more disagreed instances because we did not limit them to temporal or relative clauses only. Thus, annotator sometimes is confused whether include the clause into cue, source or content themselves or leave them to the supplements. For example, prepositions such as (*assert to the press/صرح للصحفيين*, *pointed to/أشار إلى*) in cue supplement, or in source supplement. All disagreed cases are discussed intensively with all annotators to clarify the annotation guidelines and to increase the usability of the annotation tool. Cases that are still debatable are verified by third expert who is not involved initially in this manual annotation. The fine-grained features in our pilot study increases the challenge of automatic identification of full

Text Span Agreement (agreed attribution=161)	
Cue supplement	90%
Source	97%
Source supplement	97%
Content	97%
Content supplement	96%

Table 2: inter-annotator agreement of text features using agr metric

attribution elements in short text. The pilot study results a mini-annotated corpus, Table 3 presents the annotation distribution. From the few articles (20) we found 161 instances of attribution only 60 of them are direct quotations, the rest is indirect (paraphrasing). As expected in news text 80% of the instances are declaration/assertion attribution. Supplement features are used frequently in news to present the background information related to entities and events in the augmented element. The cue used in implicit attribution are mostly indirect and influenced by the writer intention of reporting specific news. The purpose and the source of attribution may be used on validation the news and a level of reality. While there is no defined list of speech acts in Arabic for each class of purposes, we expect the disagreement will continue on this feature. We plan to study the list of cues we have in current mini corpus in terms of ambiguity and provide it to the annotators in the next phase.

Attribution instances	161		
distinct cues	85		
Explicit cue	130	**Direct At**	60
Implicit cue	31	**Indirect At**	101
Common purposes	Declaration(72), Assertion(62), Expression(14),		
Common used cues	say/qAl/قال, declare/A'gn/أعلن, add/ATaf/أضاف		
Supplements	Cue(49), source(37), content(10)		

Table 3: Pilot corpus of attribuation in Arabic

7. Conclusion

Attribution annotation in Arabic requires a comprehensive analysis of contemporary corpus due to the lack of resources in Arabic linguistics evaluating different kinds of quotation and reporting others speech and opinions. We propose annotation guidelines and annotation tool to build the first attribution corpus for MSA news articles in particular. Insight of our pilot annotation experience, the guidelines and the annotation tool are slightly adapted before conducting a full annotation of attribution in the ATB. The corpus will be a valuable resource for authorship, rumour identification, Named Entity recognition with source feature, speech acts, and sentiment and polarity systems.

8. Acknowledgment

This work is funded by the deanship of scientific research at Al-Imam Mohammad Ibn Saud Islamic University,

Saudi Arabia, No. 370904. The team is grateful to Dr. Gihan Issa, Prof. Bonnie Webber, and Dr. Fatma Alshihri for the valuable linguistic discussion.

9. Bibliographical References

Al-Saif, Amal, and Katja Markert. 2010. "The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic." In LREC.

Ali, Mohammed Mohammed Younis. 2004. Introduction to the science of significance and communication (New United Book House).

Austin, John L. 1975. 'How to do things with words (JO Urmson & M. Sbisa, Eds.)', Harvard U. Press, Cambridge, MA.

Bouayad, Nawara, and Omar Belkheer. 2012. Categorize speech actions in journalistic discourse Algerian written in Arabic, Journal of Al-Athar

Carlson, Lynn, and Daniel Marcu. 2001. Discourse tagging reference manual, ISI Technical Report ISI-TR-545, 54: 56.

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

Elson, David K, and Kathleen McKeown. 2010. "Automatic Attribution of Quoted Speech in Literary Narrative." In AAAI. Citeseer.

Elson, David K, and Kathleen R McKeown. 2009. "A tool for deep semantic encoding of narrative texts." In Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, 9-12. Association for Computational Linguistics.

Fernandes, William Paulo Ducca, Eduardo Motta, and Ruy Luiz Milidiú. 2011. "Quotation extraction for portuguese." In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011), Cuiabá, 204-08.

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

Guzmán-Cabrera, Rafael, Manuel Montes-y-Gómez, Paolo Rosso, and Luis Villasenor-Pineda. 2009. Using the Web as corpus for self-training text categorization, Information Retrieval, 12: 400-15.

Huang, Hen-Hsen, and Hsin-Hsi Chen. 2011. "Chinese Discourse Relation Recognition." In IJCNLP, 1442-46.

Ide, Nancy, and Keith Suderman. 2009. "Bridging the gaps: interoperability for GrAF, GATE, and UIMA." In Proceedings of the Third Linguistic Annotation Workshop, 27-34. Association for Computational Linguistics.

Juola, Patrick, and R Harald Baayen. 2005. A controlled-corpus experiment in authorship identification by cross-entropy, Literary and Linguistic Computing, 20: 59-67.

Kolachina, Sudheer, Rashmi Prasad, Dipti Misra Sharma, and Aravind K Joshi. 2012. "Evaluation of Discourse Relation Annotation in the Hindi Discourse Relation Bank." In LREC, 823-28.

Kong, Fang, and Guodong Zhou. 2017. A CDT-styled end-to-end Chinese discourse parser, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16: 26.

Li, Yancui, Wenhe Feng, Jing Sun, Fang Kong, and Guodong Zhou. 2014. "Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure." EMNLP, 2105-14.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. "The penn arabic treebank: Building a large-scale annotated arabic corpus." In NEMLAR conference on Arabic language resources and tools, 466-67.

Mamede, Nuno, and Pedro Chaleira. 2004. Character identification in children stories. in, Advances in natural language processing (Springer).

Mírovský, Jiří, Pavlína Jínová, and Lucie Poláková. 2014. "Discourse Relations in the Prague Dependency Treebank 3.0." In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, 34-38.

Müller, Christoph, and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2, Corpus technology and language pedagogy: New resources, new tools, new methods, 3: 197-214.

Mustafa, Mansour. 2011. The theory of verbs in the imaginary discourse between Searle and genes, Journal of Al-Athar, 12.

Neumann, Hendrik, and Martin Schnurrenberger. 2009. E-Mail authorship attribution applied to the Extended Enron Authorship Corpus (XEAC) .

O'Keefe, Tim, Silvia Pareti, James R Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 790-99. Association for Computational Linguistics.

Ouamour, S, and Halim Sayoud. 2013. "Authorship attribution of short historical arabic texts based on lexical features." In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on, 144-47. IEEE.

Pareti, Silvia. 2011. Annotating attribution relations and their features. In Proceedings of the fourth workshop on Exploiting Semantic Annotations in Information Retrieval, 19-20. ACM.

Pareti,Silvia. 2016. PARC 3.0: A Corpus of Attribution Relations, LREC.

Pareti, Silvia 2012. "A Database of Attribution Relations." In LREC, 3213-17.

Pareti, Silvia, Timothy O'Keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. In EMNLP, 989-99.

Pareti, Silvia, and Irina Prodanof. 2010. Annotating Attribution Relations: Towards an Italian Discourse Treebank. In LREC.

Pouliquen, Bruno, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news.

In Proceedings of Recent Advances in Natural Language Processing, 487-92.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Aravind K Joshi, and Bonnie L Webber. 2006. Attribution and its annotation in the Penn Discourse TreeBank, TAL, 47: 43-64.

Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual .

Rabab'ah, Abdullateef, Mahmoud Al-Ayyoub, Yaser Jararweh, and Monther Aldwairi. 2016. "Authorship Attribution of Arabic Tweets." In AICCSA

Sarmento, Luis, Sergio Nunes, and E Oliveira. 2009. "Automatic extraction of quotes and topics from news feeds." In 4th Doctoral Symposium on Informatics Engineering.

Saurí, Roser, and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality, Language Resources and Evaluation, 43: 227.

Schneider, Nathan, Rebecca Hwa, Philip Gianfortoni, Dipanjan Das, Michael Heilman, Alan Black, Frederick L Crabbe, and Noah A Smith. 2010. Visualizing topical quotations over time to understand news discourse .

Searle, John R. 1976. A classification of illocutionary acts, Language in society, 5: 1-23.

Siegel, Sidney, and NJ Castellan. 1988. Nonparametric systems for the behavioural sciences, McGraw Hill International Editions.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. "BRAT: a web-based tool for NLP-assisted text annotation." In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 102-07.

Strötgen, J. and Gertz, M. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. Proceedings of the twelve International Conference on Language Resources and Evaluation (LREC12), pages 3746–3753, Istanbul, Turkey, may.

Superman, S., Batman, B., Catwoman, C., and Spiderman, S. (2000). Superheroes experiences with books. The Phantom Editors Associates, Gotham City, 20th edition.

Wiebe, Janyce. 2002. Instructions for annotating opinions in newspaper articles .

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language, Language Resources and Evaluation, 39: 165-210.

Wolf, Florian, and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study, Computational Linguistics, 31: 249-87.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. "WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations." In ACL (Conference System Demonstrations), 1-6.

Zeyrek, Deniz, and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish Discourse Bank, LAW XI 2017: 76.

Zhao, Ying, and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In Asia Information Retrieval Symposium, 174-89. Springer.

Zhou, Lanjun, Binyang Li, Zhongyu Wei, and Kam-Fai Wong. 2014. "The CUHK Discourse TreeBank for Chinese: Annotating Explicit Discourse Connectives for the Chinese TreeBank." In LREC, 942-49.

Zhou, Yuping, and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 69-77. Association for Computational Linguistics.

Yuping Zhou Nianwen Xue. 2015. The chinese discourse treebank: a chinese corpus annotated with discourse relations, *Language Resources and Evaluation*, 49: 397-431.

10. Language Resource References

Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki, Sondos Krouna, Basma Bouziri, Wajdi Zaghouani (2010). The Penn ArabicTreeban: Part 1v 4.1. Distributed via LDC. LDC2010T13. ISLRN 512-715-458-848-0.