Abstract – This paper briefly introduces several of the aspects to take into account in order to properly describe and analyze the expression of uncertainty in textual data. Different types of ambiguity inherent to the nature of language itself are presented. Linguistic ambiguities can be observed between symbols and the meanings arbitrarily attached to them. Many natural language processing techniques can be applied to texts to minimize linguistic ambiguities. Referential ambiguities relate to the world and can be observed through extra-linguistic environments, each potentially impacting the interpretation of natural language utterance. From a linguistic point of view, the identification and automatic tagging of expressions of certainty/uncertainty in textual data is a sine qua non condition to enable the empirical study and modeling of how humans assess certainty through their use of language. Such analysis is required to generate future language-dependent models of certainty/uncertainty suitable for information fusion systems.

Keywords: Information fusion, uncertainty, ambiguity, linguistics, natural language processing, ontology.

1 Introduction

Most of the information useful to defence and security organizations in the asymmetric and counter-terrorism arena is produced by human sources and is available in the form of linguistically-based observational data. Among the several challenges related to the exploitation by computers of such sources in a fusion process, the disambiguation of word senses and the identification of the level of certainty/uncertainty linguistically expressed in those sources are key to support high-level fusion. This paper briefly introduces several of the aspects to take into account in order to properly describe and analyze the expression of uncertainty in textual data.

2 New Information Fusion Challenges

Information encoded using natural languages offers many challenges to the fusion community. As stated by Bunt [1], “computing meaning is something that we (humans) do when we read, when we write, when we listen and when we speak, and to a certain extent also when we think and when we dream. […] Computers are superior computing devices; if we could get them to compute meanings in the sense that they associate similar meanings with natural language utterances and texts as people usually do, then that would open fascinating possibilities […]”. The design of any process that could compute meaning faces the main obstacle that ambiguity, uncertainty, and vagueness are inherent to natural language.

In order for computer processes to be able to exploit linguistically-based observational data such as intelligence reports, such processes first need to disambiguate texts before any further processing. Then, once disambiguated, natural language processing (NLP) techniques should be applied to extract facts and the expression of uncertainty attached to them in the texts.

3 Certainty/Uncertainty Models

In their paper on certainty categorization, Rubin and her colleagues [2] propose a theoretical framework for manual categorization of explicit certainty information in newspaper articles. As shown in Fig. 1, the certainty markers are categorized according to four dimensions: perspective, focus, timeline, and level of certainty.

Keywords: Information fusion, uncertainty, ambiguity, linguistics, natural language processing, ontology.
characterizing information. Table 1 below shows the main categories of the proposed typology of uncertainty.

Table 1: Thomson’s typology of uncertainty

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy / Error</td>
<td>Difference between observation and reality</td>
</tr>
<tr>
<td>Precision</td>
<td>Exactness of measurement</td>
</tr>
<tr>
<td>Completeness</td>
<td>Extent to which info is comprehensive</td>
</tr>
<tr>
<td>Consistency</td>
<td>Extent to which info components agree</td>
</tr>
<tr>
<td>Lineage</td>
<td>Conduit through which info passed</td>
</tr>
<tr>
<td>Currency / Timing</td>
<td>Temporal gaps between occurrence, info collection &amp; use</td>
</tr>
<tr>
<td>Credibility</td>
<td>Reliability of info source</td>
</tr>
<tr>
<td>Subjectivity</td>
<td>Amount of interpretation or judgement included</td>
</tr>
<tr>
<td>Interrelatedness</td>
<td>Source independence from other information</td>
</tr>
</tbody>
</table>

According to Thomson et al. [3], “this typology has been developed to specifically address the types of uncertainty intelligence analysts face.” For each category of the typology, the authors also propose a set of quantitative representations. Table 2 below shows a sample of the proposed quantitative representation for uncertainty.

Table 2: Thomson’s quantitative representations (sample) [3]

<table>
<thead>
<tr>
<th>Category</th>
<th>Probability Representation</th>
<th>Parameter Examples</th>
<th>Model Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy / Error</td>
<td>Distribution of measurement error</td>
<td>Source variance $\sigma^2 = \sigma^2_0$</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>Measurement device limitations</td>
<td>Root mean square deviation $\sigma^3_j = E[ (x_j - \bar{x}_j)^2 ]$</td>
<td></td>
</tr>
</tbody>
</table>

In the fusion community, an extensive review of six different models for representing uncertainty has been performed and presented by Jousselme and colleagues in [4]. These authors promote the use of Smithson’s taxonomy of ignorance shown in Fig. 2 in support to future automated fusion systems.

4 Uncertainty in Natural Languages

It is a commonplace to state that ambiguity and vagueness are inherent to natural language. Philosophers from Plato to Wittgenstein have depicted some of the limitations of natural languages. Edward Sapir and Benjamin Lee Whorf have also provided the scientific community with linguistic determinism (e.g., language and thought are inter-dependent) and linguistic relativism (e.g., all languages do not depict the world using the same concepts).

![Figure 3: Linguistic vs extra-linguistic environments](image)

As represented in Fig. 3, humans are both in the world and in the language they use. Each human perceives the world through his/her experience and uses a shared conceptualization of this world, i.e., a natural language, to make his/her experience explicit to others. Taking into account that experiences imply three parties – the world, human “sensors”, and natural language, how can uncertainty manifest?

4.1 Dimensions of Uncertainty in Natural Languages

Since Ogden and Richards (1953), it is generally admitted that linguistically-based semiotic systems are three-fold: a symbol or “signifier” that carries a meaning or “signified” to form a linguistic sign referring to a thing or referent in the world. That is the so-called semiotic triade (Fig. 4).

Harras [5] introduces a two-level semantics dimension where the semantic form is related to the language-dependant representation of a conceptual structure and where the conceptual structure itself is related to the universal representation of encyclopaedic background knowledge, contextual information and situational conditions. This is the distinction between intra- and extra-linguistic contexts.
On one hand, different variations between a symbol and its attached meaning could generate **linguistic ambiguities**. On the other hand, different aspects such as culture could generate **referential ambiguities**, i.e., ambiguities between linguistic signs and the reality they depict in the world. Finally, natural languages have this tremendous capability to talk about themselves instead of talking about the world. This is called metalanguage:

- The concept of a *dog* does not bite.
- *Table* is a five-letter word.

### 4.1.1 Linguistic Ambiguities

Linguistic ambiguities result from many potential associations between symbols and meanings. These include polysemy and homonymy (Fig. 5). Polysemy is the capacity for a signifier to have multiple meanings or signifiers (e.g., *bank*, *table*, etc.). A homonym is one of a group of words that share the same spelling and the same pronunciation but have different meanings. (e.g. *bear*, *left*, etc.).

As described by Allen [6], each potential interpretation of a sentence, given the fact that it may contain polysemous words, corresponds to one potential logical form and to one hypothesis.ambiguous sentences will generate multiple hypothesis regarding the interpretation of the “world”. For instance the word *ball* is ambiguous. It can means “a round object used in games” or “a social event involving dancing”. Thus the sentence *Mary watched the ball* is ambiguous out of the referential context. No existing word sense disambiguation (WSD) technique (see 5.1) can disambiguate this sentence simply because the information required to determine the right meaning is outside the language realm itself, i.e., in the world (extra-linguistic information). In NLP, a single logical form can represent the two possibilities simultaneously:

\[
(\text{THE \ b1} : \{(\text{BALL1 | BALL2} \ b1)(\text{PAST (WATCH1 MARY1 b1)})\})
\]

People sharing the same natural language have to reach consensus on word and sentence meanings in order to be able to share their experience of the world. In several natural languages, such consensus is made through authoritative books such as dictionaries and grammars. Non consensual use of natural language creates idiolects. An idiolect is a variety of a language unique to a specific individual. One example of such idiolect is given by Lewis Carroll through the character of Humpty Dumpty (Fig. 6) who decides by himself what is the meaning of the words.

Other forms of linguistic ambiguities in written texts also include typos, misspellings, incomplete sentences, etc.

### 4.1.2 Referential Ambiguities

As introduced by the linguistic relativism in the 20th century, languages also convey referential ambiguities. First of all, act of speech occurs *hic et nunc* (here and now). They encode specific representations of the world as perceived by humans at a given time within a given space. During an act of speech, intra- and extra-linguistic contexts share the same space and time. As soon as the act of speech is finished, intra- and extra-linguistic contexts start drifting apart. All artefacts produced through acts of speech then become raw textual data, drifting outside their original extra-linguistic context and within new extra-linguistic contexts that might in turn impact on the original interpretation of the content.

Culture, through language, guides us in seeing and interpreting the world in terms of arbitrarily established categories (Fig. 7). As mentioned by Dennis O’Neil [7], different cultures for instance may divide up the light
spectrum in different ways. This can be seen in the comparison of some English language colors with their counterparts in the Tiv language of Nigeria:

<table>
<thead>
<tr>
<th>English</th>
<th>Tiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>green</td>
<td>(high value)</td>
</tr>
<tr>
<td>blue</td>
<td></td>
</tr>
<tr>
<td>gray</td>
<td>pupu</td>
</tr>
<tr>
<td>brown</td>
<td>nylan</td>
</tr>
<tr>
<td>red</td>
<td></td>
</tr>
<tr>
<td>yellow</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8: Different interpretations of the spectrum

As shown in Fig. 8, in the Tiv language, high value (ii) indicates light colors and low value (pupu) indicates dark colors.

At the beginning of the 20th century, Edward Sapir and Benjamin Lee Whorf studied relationships between language, culture and thought. The Sapir-Whorf hypothesis, as it came to be called, combines two principles. The first is linguistic determinism: it states that language determines the way we think. The second is linguistic relativity: it states that the distinctions encoded in one language are not necessarily found in any other language. Since, many ethno-linguistic studies have reported direct relationship between languages and the way they are used by humans to talk about their perception of their world. According to Seppo Tella [8] “in several African countries, the future lies behind us, while the past is in front of us. Think of a person rowing and you have an idea of the African notion of the tenses: while rowing in a boat, your “future” is behind your back and you need to turn your head to see where you are going. Your recent past, on the other hand, is clearly in front of you.”

One aspect of the Ojibwe pronominal system that is characteristic of Algonquian languages in general is the use of the so-called fourth person or obviative form of the third person. In essence, Ojibwe divides the category of third person into two components: the proximate third person, which refers to a more pragmatically topical antecedent and corresponds to the typical use of third person in European languages, and the obviative third person (i.e., the fourth person), which refers to a less pragmatically topical antecedent. For example, in the sentence:

Jim fooled his brother and his wife.

the reference of the expression his wife is ambiguous in English. It can refer either to Jim’s wife (the most pragmatic or the proximate meaning) or it can refer to Jim’s brother’s wife (the obviative reading). Comparable sentences are not ambiguous in this way in Ojibwe because the fourth person pronoun serves to disambiguate these meanings.

Figure 9 recapitulates the main aspects discussed above of linguistic/referential ambiguities impacting the analysis of certainty/uncertainty, and introduces some others such as intent and body language.

5 Analysis of Textual Data

Textual datasets convey interpretations about the world. They are testifying facts. The first requirement to achieve proper exploitation of linguistic data in the information fusion process is the identification and extraction of those facts from natural language texts. This requires several linguistic analyses such as tokenization, stemming, part-of-speech tagging, sentence tagging, named entities tagging, anaphora resolution, pronominal resolution, and word sense disambiguation. At the end, electronic texts must be linguistically disambiguated as much as possible. Open source frameworks such as GATE1 (Fig. 10) offer several components to perform common linguistic analysis tasks.

---

1 GATE: General Architecture for Text Engineering, University of Sheffield, UK (http://gate.ac.uk)
The second requirement is to properly assess the level of certainty/uncertainty of the facts accounted for in texts. To do so, natural language processing techniques and tools can be used to identify and tag expressions of certainty/uncertainty in texts.

Word-Sense Disambiguation

Word sense disambiguation is the task of determining the meaning of an ambiguous word within a given context. Most of work in automatic WSD is done using either supervised or unsupervised methods. Supervised methods use sense-tagged training datasets, such as those provided by the Senseval competition. Unsupervised approaches generates classifiers using very large untagged text datasets. External resources such as machine readable dictionaries, thesauri, and computational lexicons and ontologies such as WordNet, CyC and ACQUILEX can also be used in support to WSD tasks. A good state of the art has been produced in 1998 by Ide and Véronis [9]. More recent work on WSD include [10, 11, 12, 13]. Patwardhan et al. [14] describe an unsupervised WordNet-based WSD system which disambiguates a target word by using measures of semantic relatedness to find the sense of the word that is semantically most strongly related to the senses of the words in the context of the target word. Once WSD has been applied to textual data, the next task requires the identification and tagging of expressions of certainty/uncertainty.

Tagging of certainty/uncertainty expressions

Despite its linguistic complexity, certainty/uncertainty can be considered as any other knowledge domains (e.g.,

Typically, the analysis of semantic, geospatial and temporal aspects in textual data requires the use of grammars, dictionaries, gazetteers and ontologies (Fig. 11). Thus, the main task prior to any further processing consists to linguistically disambiguate the text. This task is supported by natural language processing techniques designed for word sense disambiguation.

5.1 Word-Sense Disambiguation

Word sense disambiguation is the task of determining the meaning of an ambiguous word within a given context. Most of work in automatic WSD is done using either supervised or unsupervised methods. Supervised methods use sense-tagged training datasets, such as those provided by the Senseval competition. Unsupervised approaches generates classifiers using very large untagged text datasets. External resources such as machine readable dictionaries, thesauri, and computational lexicons and ontologies such as WordNet, CyC and ACQUILEX can also be used in support to WSD tasks. A good state of the art has been produced in 1998 by Ide and Véronis [9]. More recent work on WSD include [10, 11, 12, 13]. Patwardhan et al. [14] describe an unsupervised WordNet-based WSD system which disambiguates a target word by using measures of semantic relatedness to find the sense of the word that is semantically most strongly related to the senses of the words in the context of the target word. Once WSD has been applied to textual data, the next task requires the identification and tagging of expressions of certainty/uncertainty.

5.2 Tagging of certainty/uncertainty expressions

Despite its linguistic complexity, certainty/uncertainty can be considered as any other knowledge domains (e.g.,

Typically, the analysis of semantic, geospatial and temporal aspects in textual data requires the use of grammars, dictionaries, gazetteers and ontologies (Fig. 11). Thus, the main task prior to any further processing consists to linguistically disambiguate the text. This task is supported by natural language processing techniques designed for word sense disambiguation.

5.1 Word-Sense Disambiguation

Word sense disambiguation is the task of determining the meaning of an ambiguous word within a given context. Most of work in automatic WSD is done using either supervised or unsupervised methods. Supervised methods use sense-tagged training datasets, such as those provided by the Senseval competition. Unsupervised approaches generates classifiers using very large untagged text datasets. External resources such as machine readable dictionaries, thesauri, and computational lexicons and ontologies such as WordNet, CyC and ACQUILEX can also be used in support to WSD tasks. A good state of the art has been produced in 1998 by Ide and Véronis [9]. More recent work on WSD include [10, 11, 12, 13]. Patwardhan et al. [14] describe an unsupervised WordNet-based WSD system which disambiguates a target word by using measures of semantic relatedness to find the sense of the word that is semantically most strongly related to the senses of the words in the context of the target word. Once WSD has been applied to textual data, the next task requires the identification and tagging of expressions of certainty/uncertainty.

5.2 Tagging of certainty/uncertainty expressions

Despite its linguistic complexity, certainty/uncertainty can be considered as any other knowledge domains (e.g.,

Typically, the analysis of semantic, geospatial and temporal aspects in textual data requires the use of grammars, dictionaries, gazetteers and ontologies (Fig. 11). Thus, the main task prior to any further processing consists to linguistically disambiguate the text. This task is supported by natural language processing techniques designed for word sense disambiguation.

5.1 Word-Sense Disambiguation

Word sense disambiguation is the task of determining the meaning of an ambiguous word within a given context. Most of work in automatic WSD is done using either supervised or unsupervised methods. Supervised methods use sense-tagged training datasets, such as those provided by the Senseval competition. Unsupervised approaches generates classifiers using very large untagged text datasets. External resources such as machine readable dictionaries, thesauri, and computational lexicons and ontologies such as WordNet, CyC and ACQUILEX can also be used in support to WSD tasks. A good state of the art has been produced in 1998 by Ide and Véronis [9]. More recent work on WSD include [10, 11, 12, 13]. Patwardhan et al. [14] describe an unsupervised WordNet-based WSD system which disambiguates a target word by using measures of semantic relatedness to find the sense of the word that is semantically most strongly related to the senses of the words in the context of the target word. Once WSD has been applied to textual data, the next task requires the identification and tagging of expressions of certainty/uncertainty.

5.2 Tagging of certainty/uncertainty expressions

Despite its linguistic complexity, certainty/uncertainty can be considered as any other knowledge domains (e.g.,

Typically, the analysis of semantic, geospatial and temporal aspects in textual data requires the use of grammars, dictionaries, gazetteers and ontologies (Fig. 11). Thus, the main task prior to any further processing consists to linguistically disambiguate the text. This task is supported by natural language processing techniques designed for word sense disambiguation.
terrorism, maritime surveillance, etc.). As any knowledge domains, certainty/uncertainty can be formalized in a domain ontology using well-documented ontological engineering methodologies. Some research projects currently investigate natural language processing techniques as means to support the extraction and formalization of ontologies. Among those, the SACOT project (Fig. 12) at Defence R&D Canada currently investigates NLP techniques such as contrastive terminology extraction, pattern-based semantic relations extraction, and named entities extraction to generate draft ontologies from domain-specific corpora [15, 16, 17].

One of the first steps in the formalization of a certainty/uncertainty knowledge domain is the identification of its concepts through the study of linguistic expressions. Natural languages offer several means to express uncertainty. One of the most obvious means is the lexicon of words conveying such idea of UNCERTAINTY. In natural languages such as English, the concept of UNCERTAINTY itself can be expressed using different words conveying different aspects of this idea of UNCERTAINTY. For instance, the latest edition of the Roget Thesaurus (2008) gives the following lists of synonyms and antonyms for UNCERTAINTY:

Table [3] below shows how WordNet, one of the most well-known lexical resource for English, describes the semantic network for uncertainty.

In her most recent work, Marshman [19] explores linguistic patterns used to support expressions of uncertainty in conceptual relations of ASSOCIATION and CAUSE-EFFECT. Marshman reveals several aspects in the expression of uncertainty including quantification (all, most, many, some, etc.), negative constructions, modal verbs (can, could, may, might, should, etc.), and hedges (more or less, roughly, somewhat, mostly, essentially, very, especially, exceptionally, often, almost, practically, actually, really).

All those lexical items can be used as linguistic markers to automatically analyze and tag the expressions of certainty/uncertainty attached to facts in textual information.

6 Mapping levels of uncertainty to numeric values

Since uncertainty is often seen as a classification problem in the information fusion community, many authors have tried to apply probability theory in the mapping of

---

3 SACOT: Semi-Automatic Construction of Ontologies from Texts.

4 Words whose meaning implicitly involves fuzziness [20]
qualitative assertions to quantitative values. For instance, in 1973, a report from Rand [21] mentioned previous work from Sherman Kent, an intelligence expert, to establish a mapping between the phrases used to connote various levels of uncertainty and a precise numerical range of values. Figure 14 shows one of these “Kent Charts”.

<table>
<thead>
<tr>
<th>Probability (percent)</th>
<th>Verbal Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>It is certain that ...</td>
</tr>
<tr>
<td>85-99</td>
<td>It is almost certain that ...</td>
</tr>
<tr>
<td>60-84</td>
<td>It is probable that ...</td>
</tr>
<tr>
<td>15-39</td>
<td>It is probable that ... not ...</td>
</tr>
<tr>
<td>1-14</td>
<td>It is almost certain that ... not ...</td>
</tr>
<tr>
<td>0</td>
<td>It is impossible that ...</td>
</tr>
</tbody>
</table>

Figure 14: Kent Chart of numerical values of levels of uncertainty [21]

Surveys of intelligence consumers have shown that their interpretation of the precise meaning of the words and phrases included on the Kent Chart covers an extremely wide range and is definitely not consensual.

For Druzdzel [18], the concept of subjective probability, which is the mapping of person’s belief onto the real number between 0 (event believed impossible) and 1 (event believed to occur with certainty) plays a major role in decision making. Druzdzel survey shows that people have a preference for verbal expressions instead of numerical values, reinforcing the need to develop robust linguistic analysis modules to cope with verbal expressions of uncertainty. One of the findings of the study shows that people are generally internally consistent in their use of verbal expressions of uncertainty and that there is a great between-subject variability for specific expressions like possible, probable and predictable. These same words were assigned numerical certainty values from 0.001 to 0.99 by different subjects showing a highly inconsistent interpretation of words having fuzzy semantic boundaries.

7 Conclusion

The scope of this paper was to broadly introduce most of the aspects to take into account in order to properly describe and analyze the expression of uncertainty in textual data. Different types of ambiguity inherent to the nature of language itself have been presented. Most of linguistic ambiguities can be resolved using the linguistic context (collocations, part-of-speech information, WSD, etc.). Typical resources needed to perform linguistic disambiguation include lexicons, grammars, dictionaries and algorithms. Referential ambiguities are more challenging. Their resolution requires access to extra-linguistic context and information (cultural models, time, space, mood, intent, etc.). Disambiguation of referential ambiguities requires a yet-to-come complete theory of context showing how contextual elements can trigger and actualize specific meaning values. Typical resources to support referential disambiguation would include ontologies of mental models, context models, cultural models, etc.

From a strictly linguistic point of view, the identification and automatic tagging of expressions of certainty/uncertainty in textual data is a sine qua non condition to enable the empirical study of how humans assess certainty through their use of language. Such analysis is required to generate future language-dependent models of certainty/uncertainty suitable for information fusion systems.

8 References


