Shallow semantic analysis to estimate HUMINT correlation

Valentina Dragos
ONERA-The French Aerospace Lab
Chemin de la Hunière
Palaiseau, France
valentina.dragos@onera.fr

Abstract—This paper proposes a new solution to estimate the correlation of HUMINT data. From a more general point of view, the underlying problem tackled concerns the evaluation of information, as, according to military doctrine, the more correlated data are, the more credible, therefore valuable, they should be considered. The solution proposed is completely automatic, and it is based on shallow semantic analysis. HUMINT data are not enriched by semantic annotations, as it is the case in many approaches treating such data, but they are processed by a chain of treatments allowing the identification of context specific features and ontological entities. Therefore, a measure is defined to express the correlation degree of HUMINT data by keeping the distinction between ontological entities and their evolution context. A military intelligence ontology supports this solution. It models domain entities, whose properties are exploited by shallow analysis procedures. The ontology also allows us to cope with linguistic variability, an inherent problem of HUMINT information processing. Going beyond keywords spotting and analysis, the proposed solution provides a more accurate estimation of HUMINT correlation and its output can be exploited by different applicative scenarios. Therefore the paper briefly addresses the use of shallow semantic analysis to assess uncertainty of HUMINT data.

Keywords—Information evaluation, HUMINT, shallow semantic analysis, ontology, defense and intelligence

I. INTRODUCTION

Creating effective situation awareness requires solutions able to take into account an increasing amount of data provided by different types of sources. Therefore, sensor data are supported with images, which could be at their turn enriched by annotations of human operators. By an information fusion process it becomes possible to provide an accurate picture of a given situation.

While nowadays techniques offer good results for low level information fusion, as defined by the JDL model, see [18], they prove to be limited as soon as we are dealing with higher information fusion levels, such as situation assessment, which includes identification and prediction of relations holding between entities, or impact assessment, which includes prediction of impacts or effects of current actions. Achieving high level information fusion requires new treatment paradigms, able to exploit symbolic data, which appear as a key element of this process. Symbolic data can be produced to enrich numerical ones; images are often annotated manually, for instance. They can also appear as an output of speech transcription, but open sources and HUMINT\footnote{HUMan INTelligence} are the main symbolic data sources. In spite of their rich content, those data are characterized by a large topical and linguistic variety and their exploitation inherits difficulties of treatment related to textual data.

Whether numerical or symbolic, before using a piece of information there is a need to estimate its quality. Thus, decision support relies on certain data, otherwise mechanisms are provided to keep under control the level of uncertainty. In the military field, information is evaluated depending on both the reliability of its source and its confirmation by other sources, cf. [19]. When dealing with symbolic data, the confirmation of one piece of information by another is rarely complete, and for this reason the correlation is most commonly adopted as a measure expressing the degree of overlap of two different information pieces.

The goal of this work is to develop a new method to estimate the correlation of HUMINT data, going behind keyword spotting and analysis. The proposed solution is based on shallow semantic analysis, and takes into account both the domain knowledge, in our case modeled by an ontology, and features of the applicative context. This is a non-supervised approach, that can be applied to any field for which an ontology was already created. Furthermore, outcomes provided by shallow semantic analysis provide a good basis to estimate various levels of data uncertainty, and the paper offers a brief description of this potential applicative scenario.

The outline of the paper is as follows: section 2 introduces several approaches developed to exploit textual information in the military field, with emphasis on HUMINT data. Particularities of those data, along with their underlying treatment difficulties are discussed in section 3. Section 4 details our approach to estimate HUMINT correlation; results are analyzed in section 5, which also introduces the domain ontology used in support of this work. Section 6 briefly discusses the use of shallow semantic analysis to assess information uncertainty while section 7 concludes by final remarks and perspectives for future work.
II. STATE OF ART

Human intelligence, or HUMINT, referring to gathering intelligence by the mean of human sources, is recognized as a critical capability for many military applications. It supports a broader range of operational activities, for which it has become an important asset.

Several research efforts have been conducted in order to develop effective solutions treating this rich and valuable information source. Among them, [4] describes an automated approach to classify structured military messages. The authors enrich texts by part-of-speech (POS) information and make use of rules to assign categories to messages based on POS annotations and their structure. This is a supervised solution, based on prior knowledge expressed by rules, taking advantage of the structure of messages, although more often HUMINT data are unstructured.

A solution for free-form text treatment is proposed by Hecking and colleagues, see for instance [13], [14] or [15], which are using information extraction techniques to translate natural language sentences into structured forms, called typed feature structures. Those are complex structures, highlighting different pieces of information, along with their types, that are built thanks to a chain of lexical treatments: tokenization, stemming, morphological analysis, or shallow parsing, see for instance [16]. The chain of treatments places the verb as a central element of analysis.

Rather similar to this solution, as it also places the verb as a central item, the approach described in [21] is going further by taking into account the semantic level. Thus, the authors are using an ontology of verbs, allowing them to identify verbal frames within each sentence. Verbal frames are associations between a verb and its arguments, highlighting an action along with the agents initiating, and respectively, supporting this action. While this approach is offering a semi-formal representation of free-form text, considering only verbs, on the lexical level, and therefore only roles, on the semantic level, represents its main limitation. A solution going beyond this limitation is developed by Biermann and colleagues, [3]. Authors are structuring free-form text as semantic nets, structures composed of a role and concepts connected through, taking advantages of the complexity of relations holding between ontology concepts. This representation could also be enriched by specifying the level of uncertainty.

From another interesting perspective, [20] considers both linguistic and semantic levels to create a richer description of texts. They make use of several ontologies to identify events with texts and correlated them by the mean of linguistic connectors.

A more general solution for the intelligent processing of large document collections by the joint utilisation of techniques and tools for document summarizing, semantic analysis, and classification algorithms was proposed by [18], while in [17] various natural language techniques are combined to analyse HUMINT reports.

The originality of our approach is to combine both context features and domain knowledge to express the correlation degree of HUMINT data. From an ontological point of view, this corresponds to proposing reasoning mechanisms able to keep the distinction between ontological entities and their instances.

III. PROCESSING HUMINT DATA

HUMINT data convey testimonies about relevant events, provided by human sources according to their own interpretation. When gathering data from such sources, electronic reports are created, having a structure facilitating their further exploitation. Beyond this structure, the most valuable part of those reports is the information provided by operators, hence expressed in natural language. Therefore, the underlying difficulties of HUMINT processing are related, in a more general way, to textual data. As information sources are distinct and various, collected data have a large topical and lexical variety. Moreover, linguistic data are inherently ambiguous. Further on, different types of textual data ambiguities are discussed before focusing on characteristics of HUMINT information.

A. Ambiguities of textual data

Auger and Roy, see [1], identify two types of ambiguities of textual data, based on the semiotic triad proposed by Ogden and Richards, cf. [22]. This representation corresponds to a three-fold linguistic-semiotic system, see fig. 1, expressing connections between a symbol, also called “signifier”, carrying a meaning attached to a concrete think, or object of the world.

Linguistic ambiguities (blue line, fig. 1) appear as many associations exist between symbols and their attached meanings, and they are intrinsic to natural language. Therefore, polysemy, see fig. 2, is the capacity for a signifier to have different and non related meanings, while homonyms are words having different meanings while sharing the same spelling and pronunciation, see fig. 3.
On the other hand, referential ambiguities (red line, fig. 1) are induced by socio–cultural aspects. They are the consequence of differences existing between the unique context of production of a text, including features such as space, time, author, etc., and the various possible interpretation contexts.

From another interesting perspective, we can analyze textual data according to a two-level semantics, proposed by Harras, see [12], which is defining a semantic form and a conceptual structure. While the conceptual structure refers to background knowledge, contextual information and situational conditions, the semantic form concerns the language-dependent expression of this conceptual structure. With respect to this representation, semantic ambiguities should also be considered, which appear every time a semantic form is interpreted without having enough features to recreate the complete image of the associated conceptual structure.

In the particular case of textual data, other forms of ambiguities are related to surface phenomenon, such as incomplete sentences, typos and misspellings.

The metalanguage, which is the capability of a language to express himself is not considered here, as it is irrelevant for HUMINT, a type of textual data whose particularities are discussed here after.

### B. Particularities of HUMINT data

For this work, HUMINT information corresponds to short communications of relevant events provided by human operators. The underlying applicative scenario implies a timely processing of this information, in order to support decision making at different levels of the military chain of command. Particularities of those data are declined with respect to their structure, semantic content and lexical variance.

Hence, HUMINT data are textual paragraphs composed of a small number of sentences and they report information on one unique event. Common language is used to deliver this information; therefore figures of speech such as metaphors are not employed.

The semantic content of those reports focuses on domain entities (vehicle, for instance), whose characteristics (civilian, color, type, etc.) and status (moving, etc.) are described. Sometimes, the environment could also be specified.

From a lexical point of view, the universe of discourse is more often described by the mean of named entities, and, at operator level, efforts are conducted to offer accurate information. For each operator, HUMINT is characterized by a limited employment of synonyms, and the use of terms expressing uncertainty (ex. possible). However, lexical variance occurs at global level, as many operators provide HUMINT information.

### IV. A SEMANTIC MODEL FOR HUMINT ANALYSIS

As they are intrinsically heterogeneous and non structured data, HUMINT processing is based on natural language processing techniques, supplied with formal semantics expressing the underlying knowledge of various application fields.

The goal of this paper is to propose a model to estimate the correlation of HUMINT data by adopting the principles of shallow semantic analysis (SSA). For this work, SSA is considered as defined by Ferrandez and colleagues, cf. [6], and consists in using ontologies to model domain knowledge and performing soft surface analysis of data already represented as sets of ontological entities. According to [10], an ontology is defined as a formal and explicit specification of a shared conceptualization.

There are two main advantages of using SSA to process HUMINT information. On one hand, the use of ontologies, which are artifacts modeling domain knowledge by taking into account both the conceptual and linguistic levels. The conceptual level concerns the modelling of field entities, along with relations holding between them. The linguistic level is related to the use of natural language terms to name those ontological entities. By offering this two-fold description of domain knowledge, ontologies offer a means to handle the linguistic variety.

On the other hand, performing a shallow analysis of ontological entities, based only on exploring surface relations between them, enhances capacities of formal reasoning, while maintaining a reasonable level of calculus complexity.

To apply SSA, HUMINT data should be represented as sets of ontological entities, which is carried out by the chain of pre-processing steps described here after.

### A. Pre-processing to allow SSA

Fig. 4 illustrates the overall processing cycle developed to estimate HUMINT correlations. The entry set is composed of a collection of HUMINT data, in electronic format, be that a report or a message. While SSA is providing the general framework to estimate correlations, it requires several pre-processing steps, namely: linguistic normalisation, linguistic analysis, and semantic representation of HUMINT data.
The goal of the lexical normalisation phase is to identify and remove lexical heterogeneities, which appear as the same type of information provided by different lexical forms, and concern namely: date/time expression (ex. 11 November 2011 vs. 11/11/2011); currency; geographical coordinates; metric units (ex. M vs. inch); expression of quantifiers (ex. two vehicles vs. 2 vehicles) and abbreviations (ex. poss. vs. possible). Those heterogeneities are removed by adopting a unique format to express date/time, currencies, geographical coordinates and quantifiers. Also, a referential metric system is chosen and abbreviations are replaced by corresponding words.

The output of this phase is a homogeneous corpus, employing a unique lexical format to express the same type of information, hence facilitating the linguistic analysis carried out during the next step.

Linguistic analysis concerns: identification of sentences boundaries, tokenizing, POS tagging along with identification of words stems and also named entities (NE) recognition. The output of this step is a POS annotated corpus.

Semantic representation phase is carried out in order to obtain a semantic representation of textual data. A domain ontology is used, and words of sentences are replaced by concepts and roles of which they are the linguistic expression, see tab. 1.

At the end of this phase, texts initially represented as bag of words are represented as sets of ontological entities, which allow further processing going behind the lexical level of information.

This step can be seen as the process of translating free-form text into a semantic structure, normalized with respect to an existing ontology.

### B. Estimating HUMINT correlation

Further on, we consider HUMINT information as messages $m_i,m_j$ expressed by a set of named entities $E_i,E_j$ and ontological entities $O_i,O_j$, tanks to the previously presented chain of treatments.

An aggregated correlation coefficient is defined aiming to estimate the correlation of two HUMINT messages, by taking into account two measures: the contextual similarity and the semantic similarity.

Therefore, the contextual similarity $S_c(m_i,m_j)$ is defined as the ratio between the set of named entities shared by both messages $|E_i \cap E_j|$, with respect to the overall set of named entities appearing in considered messages $|E_i| + |E_j|$.

$$S_c(m_i,m_j) = \frac{|E_i \cap E_j|}{|E_i| + |E_j|}$$

This measure expresses the similarity of HUMINT data sharing context features expressed by named entities, such as geographical location (the same city) for instance, the higher the number of shared entities, the stronger the contextual similarity between messages. Values of this measure range from 0, when there are no common entities, to 0.5, when all named entities are shared by the considered messages. With respect to ontological point of view, this measure expresses the similarity of HUMINT data at instance level, as named entities can be considered instances of ontological concepts. The underlying working hypothesis holds that named entities are on the same level of granularity.

In a similar manner, the semantic similarity $S_s(m_i,m_j)$ is defined as the ratio between the set of ontological entities, whether concepts or roles, shared by both $|O_i \cap O_j|$ messages, with respect to the overall set of ontological entities appearing in both messages, $|O_i| + |O_j|$.

$$S_s(m_i,m_j) = \frac{|O_i \cap O_j|}{|O_i| + |O_j|}$$

#### TABLE I. TEXT AS BAG OF ONCEPTS

<table>
<thead>
<tr>
<th>Concept</th>
<th>Instance of</th>
<th>Role</th>
<th>Instance of</th>
<th>Instance of</th>
<th>Status</th>
<th>Type</th>
<th>Count</th>
<th>Zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Civilian</td>
<td>Vehicle</td>
<td>4X4</td>
<td>Parked</td>
<td>Access Road</td>
<td>MSR</td>
<td>Concept</td>
<td>Concept</td>
<td>Instance</td>
</tr>
</tbody>
</table>
This measure expresses the similarity of HUMINT messages with respect to an existing ontology, which is providing a particular conceptual model of the application field. When identifying the set of ontological entities shared by both messages, the existence of equivalent relations holding between different entities is considered. Hence, two messages containing “vehicle” and, respectively “automobile”, will be considered as similar if the ontology models an equivalence relation between those concepts. Semantic similarity ranges between 0, when no common concept or role appears in both messages, to 0.5, when messages are described by the same set of ontological entities.

With respect to measures defined above, the correlation \( C(m_i, m_j) \) of HUMINT messages is defined as the sum of both contextual and semantic similarities.

\[
C(m_i, m_j) = S_c(m_i, m_j) + S_s(m_i, m_j)
\]  

(3)

The value of correlation ranges from 0, when messages are related to events occurring in completely different contexts, and involving distinct types of entities, to 1, when messages are related to, potentially, the same event.

For this work, the estimation of HUMINT correlation is carried out by analysing two types of semantic features: named entities, which can be considered as instances of concepts, and ontological entities. We consider this semantic analysis as a shallow one as only two types of relations are considered: the identity of instances, at individual level, and the equivalence of concepts and roles.

V. TEST RESULTS AND ANALYSIS

A first experimentation was conducted where the proposed solution was applied to estimate the correlation between HUMINT messages provided by a NATO tactical exercise. After introducing the ontology used to implement SSA, this section describes textual data used for this experimentation and briefly discusses the outcome.

A. ONTO-CIF ontology

For this experimentation the ONTO-CIF ontology was used. ONTO-CIF is military intelligence ontology, build in order to have a general description of this application field. It provides us with a standard model of various entities of this domain, along with the set of relations holding between them. The METHONTOLOGY methodology, see [7], was adopted to model this ontology by exploiting several knowledge supports. According to METHONTOLOGY, five steps are needed to build an ontology.

The specification step allows us to identify the purpose of the ontology construction and to explain its intended use. ONTO-CIF was created in order to describe the military intelligence field according to a functional point of view. Therefore, we used as knowledge support several documents created by domain actors with emphasis on functional descriptions of entities. The goal of the overall process was to obtain a shared standard conceptual model of our field, providing a basis for the development of further automatic reasoning mechanisms, such as evaluating information credibility. The conceptualization step identifies field entities along with their relations.

During this step are modeled domain concepts, relations between them and axioms. By exploiting knowledge sources, we have identified 58 concepts, clustered in 6 main categories, corresponding to: entities (i.e. vehicles or persons), locations (i.e. geographical area), and structure (whether plain or hierarchical, for instance), status of entities (natural or man-made), goals (or function of entities) and events (significant actions involving several entities). Fig. 5 sketches the main concepts of ONTO-CIF ontology.

![Figure 5. ONTO-CIF: main concepts](image)

We also identified 55 relations between concepts expressing: specializations (Organization, Social Grouping), compositions (Organization, Person), equivalences (Installation, Place) and field particular relations such as has-goal (Organization, Function). Axioms of ONTO-CIF highlight 4 pairs of equivalent classes, 6 pairs of disjoint classes and 4 pairs of reversed roles.

The formalization and implementation of ontology concern the choice of formalism and language to represent the ontology. We chose description logics (see [2]) and OWL DL, a sub-language of OWL (cf. [5]), as it offers a good compromise between expressiveness and computability.

The last step of the construction process is ontology maintenance and it will be done all ontology life cycle long.

B. Textual corpus

The corpus used for experimentation is composed of 81 messages reporting 6 different events, occurring during 3 days in different geographical locations.

Messages are stored as XML files whose tags provide several meta-data such as: geographical area (see fig. 6) of the reported event, time of information delivery or its security level.
As illustrated by fig. 7, time and space meta-data are used for the spatial-temporal correlation of messages, therefore the collection of messages is divided into several classes. Each of these classes contains messages whose information is delivered almost simultaneously, and it concerns events occurring in the same geographical area, although they can not be correlated.

For this experimentation 27 classes, having between 2 and 26 messages have being generated by the spatio-temporal correlation. The proposed approach to estimate HUMINT correlation is applied at class level.

C. Results analysis and discussion

For this experimentation, we do not propose a formal evaluation of the outcome, as it is difficult to have a “ground truth” of data used. Instead of this, obtained results are analysed in order to identify particular data cases and limitations of the proposed solution. For following examples, values of the correlation coefficient appear in parenthesis. Hence, this analysis highlights several cases which are representatives for the considered textual corpus, when significant correlation values are assigned to messages sharing:

1. only named entities (fig. 8);
2. only ontological entities (fig. 9);
3. named entities and domain concepts (fig. 10).

Confirmation of messages, which is a particular case of correlation, is illustrated in fig. 11, while fig. 12 shows independent messages:

- **Confirmation of messages**
  - 2 DOG EAR RADAR (100 %)
  - DOG EAR RADAR (100 %)

- **Independent messages**
  - 2 DOG EAR RADAR (0 %)
  - 2 Suspect MLRS systems heading south-west (0 %)

Limitations of the proposed approach appear as a consequence of performing only surface analysis of ontological entities, ignoring deeper relations holding between them such as subsumption or disjunction of concepts or the existence of reversed roles.

Hence, in fig. 13, messages are assigned an important correlation score, although they provide contradictory information as concepts “east” and “west” are disjoint.

In a similar manner, messages in fig. 14 are expressing contradictory actions.

- **Disjoint concepts are ignored**
  - Unknown vehicles moving east (22 %)
  - Convoy (16 vehicles) moving west direction (22 %)

- **Reversed roles are ignored**
  - Convoy passing dog ear (25 %)
  - Convoy stopping at site (25 %)

From a linguistic point of view, processing negative sentences also appears as a limitation of our approach, as for now negations are not taken into account.

To improve this solution and validate its pertinence, more experimentations are needed, exploiting larger textual corpora from various application scenarios.

VI. ON USING SHALLOW SEMANTIC ANALYSIS TO ASSESS UNCERTAINTY

Uncertainty is a fundamental aspect of information and underlines much of its treatments, becoming a crucial issue when impacting decision support tasks. The model proposed to estimate HUMINT correlation is also able to provide insights about uncertainty, as, according to the military point of view, the more correlated two information pieces are, the more credible, the more certain, they should be considered. More particularity, this model can be used or extended to assess three specific criteria defined by the uncertainty representation and reasoning evaluation framework (URREF2).

Those criteria are vagueness, inconsistency and self-confidence.

---

2 http://eturwg.c4i.gmu.edu/
Hence, the URREF defines *vagueness* as a measure of how vague, understood as not precise or inaccurate, a piece of information is. Our model can be used to estimate the vagueness of an information item as it requires the representations of this item as a set of context specific elements (expressed by named entities) and domain specific entities (expressed by roles and concepts). Therefore, the vagueness can be related to the absence of contextual elements and the presence of high level concepts. The following example illustrates various degrees of information vagueness. Hence, the first sentence (S1) offers some imprecise information about the presence of persons within some area, while the second one (S2) is reducing the vagueness by specifying the status on entities, their current action and a more precise location. The last sentence (S3) offers a concrete piece of information, as the location, the activity and the entity itself are indicated.

\[
\begin{align*}
\text{(S1)} & \quad \text{Two persons on the road in the construction area} \\
\text{(S2)} & \quad \text{Civilian 4x4 parked on access road to MSR} \\
\text{(S3)} & \quad \text{In NAI 31, new detection of DOG EAR radar}
\end{align*}
\]

**Figure 15.** Various degrees of informations vagueness

The *inconsistency* is defined by URREF as a measure of the extent to which information is explicitly contradictory or conflicting. With respect to this definition, our model allows to express the inconsistency of two information pieces by taking into account the ration of disjoint classes and reversed roles occurring in their respective representations. The example below illustrates a conflictual situation, where two sentences provide contradictory information about entity evolution.

\[
\begin{align*}
\text{Convoy stopping at site} \\
\text{Convoy (16 vehicles) moving west direction}
\end{align*}
\]

**Figure 16.** Conflictual information

It is also possible to extend the proposed model in order to estimate the *self-confidence*. This a measure of the information credibility as estimated by the source itself, expressing the thrust accorded by author to the information it provides. In order to estimate this measure, the model should be extended by enriching the chain of linguistic treatments with functionalities able to identify linguistic markers expressing the certainty or uncertainty level of the information. Such markers are considered as meta-data of the information conveyed and they can be used as a basis to estimate the self-confidence. As shown here-after, some markers translate author’s beliefs (possible); others are related to its knowledge (unknown).

\[
\begin{align*}
\text{Possible hostile MLRS} \\
\text{unknown vehicles moving east}
\end{align*}
\]

**Figure 17.** Linguistic markers of uncertainty

Information evaluation process can be improved by taking into account values of those measures at different processing stages. Hence, we can avoid taking into account vague information. It becomes also possible to prevent the treatment of doubtful information, having a low value of self-confidence, and finally, this allows us to highlight inconsistent information pieces, by pointing out contradictions within the data sets under analysis.

VII. CONCLUSION AND FUTURE WORK

HUMINT data are valuable resources to achieve different goals of higher level information fusion, but processing such data appears as a difficult task due to the absence of structure and their intrinsic heterogeneity. Assessing the quality of information is a crucial step in a military context, as credible information should feed the basic building blocks of decision support systems.

In this paper we propose a semantic approach to estimate the correlation of HUMINT data. The underlying working hypothesis states that an important correlation degree corresponds to a higher level of credibility. Our model computes the correlation degree by taking into account both context features and domain knowledge, while the overall process is supported by a domain ontology. First, this ontology facilitates the translation of free-form texts as sets of ontological entities, and then it provides a basis for enhanced reasoning mechanisms.

Compared to other approaches, we do not enrich the initial data by semantic annotations, but rather keep the distinction between features expressed at entity level, thus anchoring the analysed report into a particular context, and features related to ontological knowledge, taking advantages of more general inference mechanisms.

The first direction to improve this approach is to carry out the transition form a shallow to a deep semantic analysis, by exploiting the richness of different types of ontological relations. This could lead to the definition of new measures of similarity, taking into account not only commonalities, but also differences between HUMINT information.

Another future work direction concerns the extension of this approach by computing several quality measures such as the coherence coefficient mentioned by Gurevych and colleagues see [9], and several evaluation criteria defined by URREF. This could improve the overall process, by avoiding taking into consideration deprecated, inaccurate or inconsistent information.

REFERENCES


[19] [NATO, 2003] NATO STANAG 2511 Intelligence Reports, 2003

