A Rough Set Formalization of Quantitative Evaluation with Ambiguity

Patrick Paroubek* **and Xavier Tannier*†**

*LIMSI-CNRS, 91403 Orsay, France
†Univ. Paris-Sud, 91403 Orsay, France
{pap,xtannier}@limsi.fr

Abstract

In this paper, we present the founding elements of a formal model of the evaluation paradigm in natural language processing. We propose an abstract model of objective quantitative evaluation based on rough sets, as well as the notion of potential performance space for describing the performance variations corresponding to the ambiguity present in hypothesis data produced by a computer program, when comparing it to the reference data created by humans. A formal model of the evaluation paradigm will be useful for comparing evaluations protocols, investigating evaluation constraint relaxation and getting a better understanding of the evaluation paradigm, provided it is general enough to be able to represent any natural language processing task.

Keywords: evaluation, formalization, rough sets

1. A Set Model of Objective Quantitative Evaluation

To our knowledge, no formal framework exists for studying the evaluation paradigm; we propose to lay the foundation for such model based on the mathematical notion of "rough sets"(Skowron et al., 2002), particularly adapted for reasoning about ambiguity in natural language. Natural language is ambiguous for a large part and an extra amount of ambiguity is brought by the evaluation paradigm where, a reference data, created by humans, is compared to hypothesis data, yielded by a computer program.

We propose to consider the notion of potential performance space, for describing the performance variations corresponding to the ambiguity present in the hypothesis data. A formal model of the evaluation paradigm will be useful for comparing evaluations protocols, investigating evaluation constraint relaxation and getting a better understanding of the evaluation paradigm, provided it is general enough to be able to represent any natural language processing task. With the evaluation paradigm, a computer output is compared to a reference dataset of human origin. The underlying hypothesis is that similarity of the two datasets of sufficiently large size should be considered as a proof that the computer computer model is a faithfull emulation of the human processing.

In our terminology, the "control task" is the information processing task whose performance we wish to assess, when done by some computer system. For instance in Information Retrieval (IR), given a set of documents and a query, the control task is the identification of the documents which are relevant with respect to the given query.

Control tasks can vary greatly in nature. Identifying objects or classes of objects present in the input data is one of the most straighforward control task, *e.g.* POS tagging (Paroubek, 2007) or named entity recognition (Nadeau and Sekine, 2007). In some cases, the instances of objects of interest are identified and the systems has only to identify the class they belong to (*e.g.* IR or Word Sense Disambiguation (Edmonds and Kilgarriff, 2002)). But the control task can also be much more complex in nature, for instance when the aim is to identify objects, primitive relations holding between objects and higher level relations holding between primitive relations, like for parsing (de la Clergerie et al., 2008), anaphoric resolution (Vilain et al., 1995) or image recognition (Unnikrishnan et al., 2007). In other cases, we are only interested by the final product of the transformation of the objects and relations identified in the input data, like in Machine Translation.

A control task is thus a process that links test data units produced by a segmentation of the input data to output data units, possibly organized in a hierarchy. Output units can be seen as annotating the input units, which caused their creation what ever their nature, *e.g.* in a translation task from French to English, "The" can be considered as annotating "Le" in the sentence "Le restaurant est fermé" (*The restaurant is closed.*) and in a parsing task, the syntactic dependencies can be considered as annotating the input sentence. Assuming that the test data is the result of a segmentation process of an input medium (character stream, speech signal, etc.) represented by $S = \{s_i/0 \leq i \leq N \in \mathbb{N}\}$, and the set of annotation labels by A, the m layers of relations graphs ρ resulting from the annotation process can be expressed as follows[1]:

$$\begin{aligned}
\rho &= \bigcup_{j=1}^{m} \rho_j, \quad m \in \mathbb{N} \\
\rho_1 &= \bigcup_{k=1}^{q} \{r_l/l \in \mathbb{N}, r_l \subset \mathcal{P}(S^k \times A)\} \\
\rho_i &= \bigcup_{k=1}^{u} \{r \subset \mathcal{P}((S \cup \rho_{x_1}) \times (S \cup \rho_{x_2}) \\
&\qquad \cdots \times (S \cup \rho_{x_k}) \times A), 1 \leq x_k < i\}
\end{aligned} \tag{1}$$

ρ_0 represents the first layer of annotation of the test data and ρ_i the successive layers of annotations, which can address other annotations from any layer.

Note that while the *annotation graph* model from LDC (Bird and Liberman, 2000) encodes the direct relationship between annotations and events from the various linear input streams, we represent in our model the potentially recursive structure of the annotations.

[1] In formula 1, $\mathcal{P}(x)$ is the set of all subsets of x.

For quantitative evaluation, when comparing the set of relations identified by the computer H (*hypothesis* data) with the one identified by humans R (*reference* data), the performance result is obtained by computing some *measure* defined over the two previous sets of relations. The result of evaluation, a measure $\rho \times \rho \rightarrow \mathbb{R}$, is a function of S the segmented test data, whose role in linking reference and hypothesis data is essential for the computation of the evaluation result. It may happen that the system under test uses relations ρ', whose semantics differs from the one of the reference (but nervertheless remains mappable to). It may also happen that the input data is modified by the tested system because of noise, data corruption or specific normalization, or that the segmentation function used by the system is different from one that was used to process the reference data, or both. With S' the new segmentation function, the hypothesis can then be better described by :

$$\begin{aligned} H = & \bigcup_{j=1}^{n} H_j, n \in \mathbb{N} \\ H_1 = & \bigcup_{k=1}^{q} \{S' \neq S, A' \neq A, r_l \subset \mathcal{P}(S')^k \times A')\} \\ H_i = & \bigcup_{k=1}^{u} \{r \subset \mathcal{P}((S' \cup H_{x_1}) \cdots \times (S' \cup H_{x_k}) \\ & \times A'), x_k < i\} \end{aligned} \tag{2}$$

In addition to the mapping μ from relation annotation labels A' provided by the system to the reference labels A, one then must be able to find a "reasonable" mapping M between the segmentations S and S' to be able to compute an evaluation result. Here, reasonable means a mapping that maximizes the global similarity between the reference and hypothesis data with respect to a particular similarity function σ, *e.g.* using dynamic programming to find the mapping that minimizes the edit distance between two slightly different versions of the same text to compare their POS tag annotations (Paroubek et al., 1998).

$$\begin{aligned} & M = \arg\max \textstyle\sum_{h,g} \sigma(m(h), g), \\ & m \in \mathcal{P}(S') \rightarrow \overset{m}{\mathcal{P}}(S), \\ & \sigma : \mathcal{P}(S) \times \mathcal{P}(S) \rightarrow [0, 1] \end{aligned} \tag{3}$$

Often in an evaluation campaign, the organizers define several measures in conjunction and use the vector space corresponding to the measurement tuples to synthesize the performance comparison with euclidian distance.

2. Enumerating Events

A quantitative objective evaluation results is a function of the assessment of the relative similarity (/dissimilarity) between the annotations produced by the system under test and the gold standard. In the most general case, a similarity measure is a function of the three subsets: $TP = R \cap H$, $FN = R \backslash (R \cap H)$ and $FP = H \backslash (R \cap H)$, true positive, false negative and false positive annotations (Manning and Schütze, 2002). As measure we often use: "accuracy"(Labatut and Cherifi, 2011), "error" (eq. 4), the Jaccard coefficient (eq. 5), or the F-measure (eq. 6) which combines precision and recall.

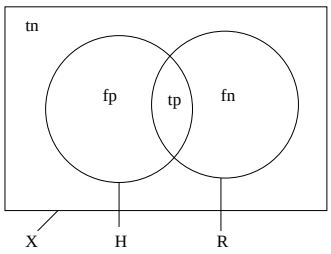

Figure 1: Subset support of quantitative evaluation (Manning and Schütze, 2002).

$$a = \frac{|H \cap R| \cup (X \backslash |H \cup R|)}{|X|}, e = \frac{|(H \cup R) \backslash (H \cap R)|}{|X|} \tag{4}$$

$$j = \frac{|H \cap R}{|H \cup R|} \tag{5}$$

$$f = \frac{1}{\frac{\alpha}{p} + \frac{(1-\alpha)}{r}}, 0 < \alpha < 1, p = \frac{|H \cap R|}{|H|}, r = \frac{|H \cap R|}{|R|} \tag{6}$$

In the end, independently of the evaluation protocol and its annotations, all performance measures are function of the cardinal of the subsets: TP,FP,FN and TN.

3. A Rough Set Model of Ambiguity

As we have seen, the results depends on the criteria for deciding whether an annotated item from the gold standard is the same as the corresponding annotated item of the hypothesis data, i.e. on the equality relations defined over the ρ_i (see eq. 1). Sometimes a strict equality relation is considered a too strict criteria and one would prefer to have more a gradual decision function, considering several answers as acceptable, possibly with different degree of acceptability. As evaluator, we can :

- authorize the systems under test to give alternative annotations for a test item instead of a single one (Resnik and Yarowsky, 2000), *i.e.* use an equivalence relation instead of an equality relation. Its classes are defined by extension in the reference data.
- incorporate in the evaluation protocol an equivalence relation instead of an equality relation (Paroubek et al., 2006).

Taking into account ambiguity in our model requires to shift from a classical set theory model to the *rough set* model (Skowron et al., 2002) of Zdzisław Pawlak (Pawlak, 1982), which fits perfectly the situation of hypothesis (and reference) annotations with ambiguity (see Figure 2). In the classical set theory, the boundaries between sets are *crisp*, *i.e.* there is a clear cut distinction between its inside and its outside. In a rough set, there is a boundary region between the two, made of the elements that could belong to the set under certain conditions. The boundary region of a rough set is made of the elements that validate some of the predicates defining the inside, but not all the predicates. The elements that do not validate any of these predicates constitute the outside. Rough sets make explicit the granularity of

information associated to the definition of a set [2]. A rough set can be numerically characterized with the *accuracy of approximation* coefficient (Komorowski et al., 1999).
Let $\mathcal{A} = (U; A)$ be an *information system* (*i.e.* a subset of $U \times A$) and let $B \subseteq A$ and $X \subseteq U$. X can be approximated with the information contained in B, by constructing the B-lower and B-upper approximations of X : $\underline{B}X$ and $\overline{B}X$ respectively, where $\underline{B}X = \{x/[x]_B \subseteq X\}$ and $\overline{B}X = \{x/[x]_B \cap X \neq \emptyset\}$[3]. With this notations the accuracy approximation coefficient is:

$$\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|} \tag{7}$$

H and R defined in section 1. are *information systems*. In the evaluation process, if we consider an equivalence relation $\approx$ instead of of an equality relation, H (and R if it contains ambiguous annotations) approximate the theoretical hypothesis (and reference) sets with respect to the equivalence criteria $\approx$. The accuracy approximation coefficient can serve to quantify the amount of change induced in the performance space by using an equivalence relation instead of an equality relation (see section 3.2.). For instance, among all the previous measures of section 2., if we consider precision p (cf eq. 6), its value $p_{\approx}$ will be as follows:

$$p.(1 - \alpha_{\approx}(H)) \leq p_{\approx} \leq p.(1 + \alpha_{\approx}(H)) \tag{8}$$

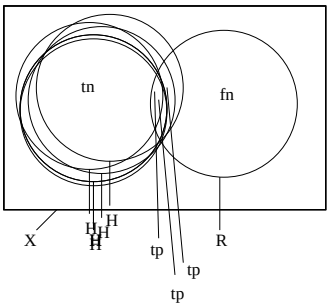

Figure 2: Rough sets and quantitative evaluation with ambiguity.

3.1. Measuring Ambiguity

A corollary to introducing ambiguity in the evaluation protocol is the wish to gauge how much a systems uses this possibility to improve its performance value. One will want to know whether the system has clearly identified problematic annotation cases, difficult even for human annotators or whether its simply tried to improve its chances at finding a correct answer by providing a larger number of answers. This can be evaluated by the measure of *decision*, proposed (Paroubek et al., 1998) to assess the level of annotation disambiguation, by measuring the average number of tokens completely disambiguated. The resulting value is located in the interval $[0, 1]$ and provides a uniform measurement for both complete and partial tagging disambiguation schemas (Pak and Paroubek, 2010).
In more general terms, *decision* D is then the ratio between the number of all the equivalence classes of size 1, over the total number of equivalence classes defined by the equivalence relation over the annotations.

$$D = \frac{|\{x/|[x]_{\approx}| = 1\}|}{|H/_{\approx}|} = \frac{|\{x/[x]_{\approx} = \{x\}\}|}{|H/_{\approx}|}$$

The *decision* measure will give us the means to quantify how far we are from a fully deterministic system.

3.2. The Potential Performance Space

Relaxing the task constraints or modifying the reward function (or both) necessarily leads to modification of the measurements takens, and if ambiguity annotation is allowed, the hypothesis data may contain ambiguous annotations, it is then legitimate to ask oneself, what would have become of the performance of the considered systems, if it had attempted full disambiguation, what is the limit performance range defined by failures or successes at disambiguating the remaining (partially) undecided annotations. In that context, the possible variation range of the evaluation parameters defines what we call the *potential performance space*. The measure space S_M is the structure formed by the collection of vectors constituted by the different measures of an evaluation process. For example, the combined precision/recall values form a two-dimensional euclidian space. Inside this space S_M, the potential performance space PPS, is the subspace formed by the collection of possibly accessible vectors, with respect to the given evaluation protocol defined by its equivalence relation.
The amount of variability in performance associated with partially disambiguated hypothesis data can be quantified with the accuracy approximation coefficient defined in formula 7 of section 3. for rough sets. The set of true positive hypothesis items is then approximated by its lower and upper approximations as defined by the reference items and the ambiguity resolution applied to ambiguous hypothesis items (see figure 2).

$$\alpha_R(tp) = \frac{|\underline{R}tp|}{|\overline{R}tp|} \tag{9}$$

For evaluations where the order of annotation units is important, and/or where decision on some units has consequences on choices concerning other units, the "decision" dimension (or "time" as decision order dimension) is important and should be considered through the different possible ways to explore the potential performance space. Figure 3 shows the evolution of the PPS of a precision measure on a basic annotation task, as decisions are made. PPS can have very different slopes when annotations are not independent from each other. In this case, making a single decision can lead to more or less choices about further dependent annotations, thus distending the potential performance space. Considering this potential performance space is interesting for the evaluator because it sheds some light on the precision of the evaluation measure itself and it is interesting for the participant, because it can provide hints at

[2] See http://chc60.fgcu.edu/EN/HistoryDetail.aspx?c=12 for history and tutorials on rough sets

[3] $[x]_B$ denotes the equivalence class of the B-indicernibility relation for element x, *i.e.* the subset of X made of all the elements indiscernible from x according to the attribute set B.

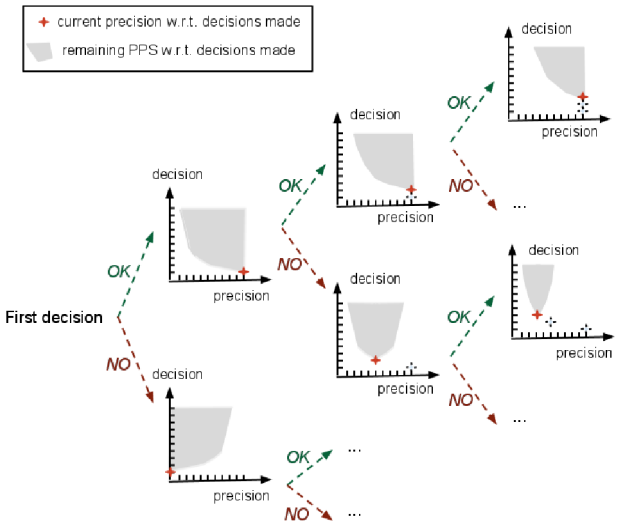

Figure 3: Evolution of Potential Performance Space (PPS) of precision measure according to decisions.

potential performance limitations or improvements. In particular, information about the boundaries of the space, what is the best or worse performance, reachable from a given measure point if ambiguity would be fully resolved.

Reward function Depending of its aims, an evaluation can be more or less strongly related to a particular applicative context. As a results, it is sometimes needed to bias the performance measures to take into account some specificity of the applicative context, for instance privileging recall over precision in security-oriented information retrieval by acting on the value of the β parameter of the F-measure. This is a notion of *reward function* in the process of evaluation performance computation. While the equality or equivalence relation defined by the protocol tells us which items are correct annotation items, the reward function tells us what bonus do we get by finding the correct annotation, or finding an annotation that is a "reasonable" approximation of the correct one. Both modifying the equality relation or the reward function has an impact on the performances measured, but the implications are different. While the *equality function* is a technology-oriented (intrinsic evaluation), the reward function is a user-oriented (extrinsic evaluation). It has no theoretical link with the operational semantics of the control task or its representation. Changing the reward function does not change the comparisons results between reference and hypothesis data.

4. Applying the Model to Real Evaluations

In this section, we show that our model can represent well known evaluation protocols from different domains of natural language processing (Paroubek et al., 2007).

4.1. Classification

For this kind of task, the purpose is to segment a data set in order to highlight parts of this set that belong to specific classes (predefined or not), and possibly to provide relations existing between these parts.

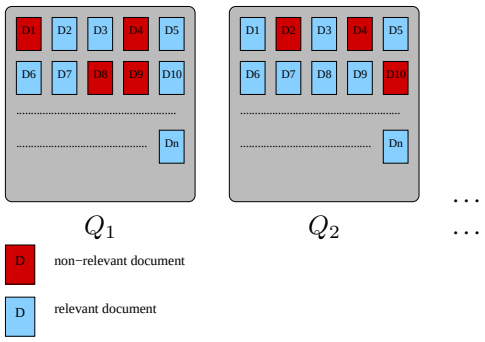

Figure 4: Classification of the set of documents among two parts (relevant and non relevant) for each query: for Q_1, D1, D4, D8 and D9 are relevant; for Q_2, D2, D4 and D10 are relevant.

We take the examples of information retrieval, named entity recognition, temporal annotation and parsing.

Information retrieval Classical information retrieval aims at finding full documents that are relevant to a given query Q. The document collection C is the input data (the retrieval unit is the entire document level). Here we consider a simplified instance of the general model presented in the previous section, in a sense that the segmentation of the test data into units to be annotated is provided, it is made of the documents themselves, see Figure 4.
This is a classification task, since the aim is to produce a partition of the collection, between relevant and non-relevant documents, with respect to the query. In practice the evaluation data contains several queries, but since in general they are considered independent of each other, the evaluation resolves to a series of single query evaluation.
In other words, variables introduced in Section 1. are instanciated in the following way:

S : the structuration corresponds to the existing document boundaries,
A : a set of two labels: *relevant for the query* or *not relevant for the query*
$\rho = \rho_1$: a singleton made of one unary relation that tag the relevance of each documents.

Named entity recognition The following steps can be identified concerning named entity recognition:

1. Identification: finding which data units of the test set need to be annotated.
2. Categorization: finding the appropriate relation to annotate a data unit from the test set, *e.g.* tagging word sequences with labels for locations, persons, organisations, etc.

A normalization step can be added, as for example at the Temporal Expression Recognition and Normalization (TERN) Task of EVALITA (Magnini et al., 2008), where temporal expression should be associated with a universal representation of the expression. All NE types can be concerned by this normalization, for example person names, since they exhibit often many variations in their realization: "Barack Obama", "B. Obama", "President Obama", "Barack H. Obama".

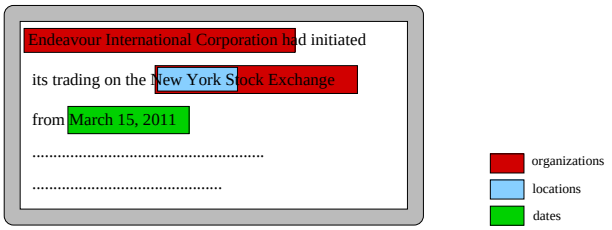

Figure 5: Classification of the sets of characters considering the named entity types (here, organizations, locations, dates, none).

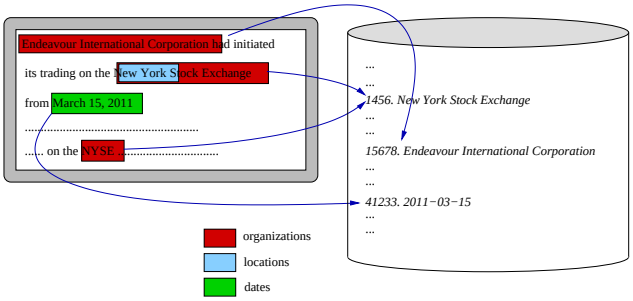

Figure 6: Classification of the sets of characters considering the named entity types and relations to normalized entities in a separate knowledge base.

Note that for named entity recognition, the segmentation function of the test data into elementary units is generally not provided by the evaluation organizers. This is not the case for the following example: TempEval.

S : the segmentation of NE types, at character or word level

A : the set of NE class labels

$\rho = \rho_1$: a singleton holding the unary relation linking the NE to its class label.

Temporal annotation Temporal annotation as defined by TempEval evaluation campaign (Verhagen et al., 2007) consists in the following: given a set of test texts for which sentence boundaries are annotated, as well as all temporal expressions and events in texts, the control task goal is to link events to other events, or events to time expressions (see Figure 7).

S : the segmentation of the token stream into temporal expressions, signals, and events.

A : the set of temporal expression signal and event class labels, as well as temporal relations labels.

$\rho = \rho_1$: 1/ the relation that links a temporal expression, signal or event to its class label, *e.g.* *kipnapped* is an event.
2/ plus all the labeled time relations between the temporal elements *e.g.* *kidnapped* is **before** *rescued*.

Parsing The aim of automatic parsing is to provide a complete/partial structural analysis of a sentence expressed in terms of:

- chunks, sequences of words with some syntactic meaning,

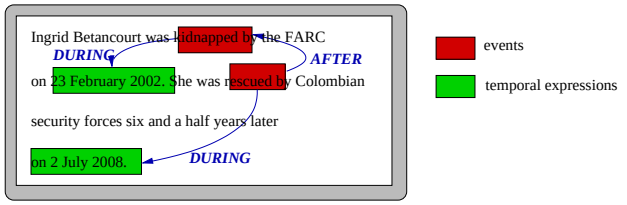

Figure 7: Temporal annotation.

- constituents, sequences of words which function as a single units within a hierarchical structure,
- dependencies, relations linking a particular word (the head) and one of its dependents,
- links, relation between pairs of words whithout necessarily referring to a tree hierarchy,
- grammatical relations, *i.e.* relation/head/dependent tuples (Watson et al., 2005),
- derivation/derived tree (Schmitz and Le Roux, 2008) describing the construction of the syntactic parse tree,
- etc.

Since theories and annotation schemes are quite numerous and diverse in parsing, we present here only a few annotation schemes which have been used for evaluation: the PennTreebank (Marcus et al., 1993) for constituent analysis of English and PASSAGE (Vilnat et al., 2010) for chunks and grammatical relations in French. The PennTreebank example is the first example of annotation scheme in this article which exhibits both relations between annotated elements (words) and their class label (*e.g.* the relation between NP-SBJ and "I"), as well as relations between annotations themselves (*e.g.* the toplevel relation between S and the constituents NP-SBJ and VP).

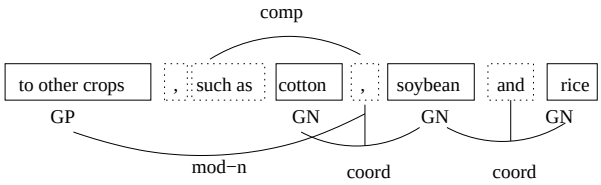

Figure 8: Example of PASSAGE annotation

The Penn Treebank constituent annotation model:

S : the segmentation into words units of the input stream

A : the set of constituent labels

ρ_1 : relations between words and their deepest layer of constituent label

ρ_i : relations between words and constituent labels, or between constituent labels of deeper levels

ρ : $\bigcup_{j=1}^{m} \rho_j, m \in \mathbb{N}$

The PASSAGE annotation scheme has only one layer of non-recursive syntactic chunks and grammatical relations defined between words and/or chunks (cf Figure 8).

Element of comparison between the passage annotation scheme and PARC, SD and GR, three other syntactic annotation schemes used for English parsing evaluation are provided in (Paroubek et al., 2009).

The PASSAGE annotation model:

- S : the segmentation into words units of the input stream
- A : the set of chunk and relation labels
- ρ_1 : the relations between words and their chunk labels, or relations for which at least one argument is word (*e.g.* coordinating relation for whose coodinating conjunction argument is always a single word not included in any chunk, see Figure 8)
- ρ_2 : the relations linking chunks only
- ρ : $\rho_1 \cup \rho_2$

4.2. Transduction

Lastly, a very different type of applications is the set of applications producing an output that is not an enrichment (or annotation) of an existing test set, but a new object obtained by transformation from or in response to another object. Examples of transduction applications are: machine translation, speech synthesis, automatic summarization, language generation (Koller et al., 2010) or machine dialogue.

For all these examples, T is a document seen as a sequence of characters to be either translated, synthetized, summarized, etc. S is the existing segmentation into language units, while A is the result of the operation: the translation of a language unit into the target language, the synthesis of a language unit into sound generation instructions, etc. R is the set of links between language units in the test set and elements from A.

- S : the segmentation of the input stream into transduction source units
- A : the corresponding transduction target units labels
- $\rho = \rho_1$: relations between source and target units

Note that we can consider multilingual alignment tasks (chuang Chiao et al., 2006) to be degenerate cases of transduction task, where the target labels are provided as input data and the systems under test need only to identify the relation between source and target units.

5. Conclusion

We have first proposed a model of objective quantitative evaluation for natural language processing based on rough sets, distinguishing the annotation equivalence relation (intrinsic criteria) and the reward function (extrinsic criteria), second we presented the notion of potential performance space to describe the effect that resolving the remaining ambiguity of the hypothesis data would have on the performance range. We have also shown that the accuracy approximation coefficient used to quantifies the level of "roughness" of a rough set can be used to describe the amount of variability of the potential performance space corresponding to the ambiguity present in the hypothesis data. Our future work will concern using and refining our model in order to obtain, from the formal representation, results that help compare, design and validate evaluation protocols.

6. References

Steven Bird and Mark Liberman. 2000. A formal framework for linguistic annotation. *Speech Communication*, 33:23–60.

Yun chuang Chiao, Olivier Kraif, Dominique Laurent, Thi Minh, Huyen Nguyen, Nasredine Semmar, Franois Stuck, Jean Véronis, and Wajdi Zaghouani. 2006. Evaluation of multilingual text alignment systems: the arcade ii project. In *Proceedings of the Fifth LREC*, pages 1975–1979, Genoa, Italy, may.

Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Nat. Lang. Eng.*, 8:279–291, December.

Alexander Koller, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2010. Report on the second nlg challenge on generating instructions in virtual environments (give-2). In *Proceedings of the International Natural Language Generation Conference (INLG)*, Dublin.

Jan Komorowski, Lech Polkowski, and Andrzej Skowron. 1999. Rough sets: A tutorial. In *Lecture Notes of the 11th European Summer School in Logic Language and Information (ESSLLI)*. http://folli.loria.fr/cds/1999/esslli99/courses/skowron.html.

Vincent Labatut and Hocine Cherifi. 2011. Accuracy measures for the comparison of classifiers. In *Proceedings of the 5th International Conference on Information Technology*, Amman, Jordan, May.

Bernardo Magnini, Amedeo Cappelli, Fabio Tamburini, Cristina Bosco, Alessandro Mazzei, Vincenzo Lombardo, Francesca Bertagna, Nicoletta Calzolari, Antonio Toral, Valentina Bartalesi Lenzi, Rachele Sprugnoli, and Manuela Speranza. 2008. Evaluation of natural language tools for italian: Evalita 2007. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth LREC*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

C. D. Manning and H. Schütze. 2002. *Foundation of Statistical Natural Language Processing*. Massachusetts institute of Technology Press, 5ème edition.

M. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19.

E. de la Clergerie, O. Hamon, D. Mostefa, C. Ayache, P. Paroubek, and A. Vilnat. 2008. Passage: from french parser evaluation to large sized treebank. In ELRA, editor, *In proceedings of the sixth LREC*, Marrakech, Morroco, May.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Journal of Linguisticae Investigationes*, 30(1). URL http://nlp.cs.nyu.edu/sekine/papers/li07.pdf.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh LREC*, pages 1320–1326, Valletta, Malta, may. European Language Resources Association (ELRA).

Patrick Paroubek, Josette Lecomte, Gilles Adda, Joseph Mariani, and Martin Rajman. 1998. The grace french part-of-speech tagging evaluation task. In *Proceedings of the First LREC*, pages 433–441, Granada, Spain, May. ELDA.

Patrick Paroubek, Isabelle Robba, Anne Vilnat, and Christelle Ayache. 2006. Data, annotations and measures in easy - the evaluation campaign for parsers of french. In *proceedings of the fifth LREC*, pages 315–320, Genoa, Italy, May. ELRA.

Patrick Paroubek, Stéphane Chaudiron, and Lynette Hirschman. 2007. Principles of evaluation in natural language processing. *Traitement Automatique des Langues (TAL)*, 48(1):7–31.

Patrick Paroubek, Eric de la Clergerie, Sylvain Loiseau, Anne Vilnat, and Gil Francopoulo. 2009. The passage syntactic representation. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*, pages 91–102, Gröningen, January. Netherlands Graduate Schools of Linguistics (LOT).

Patrick Paroubek, 2007. *Evaluation of Text and Speech Systems*, volume 36 of *Text, Speech and Language Technology*, chapter Evaluating Part Of Speech Tagging and Parsing, pages 97–116. Kluwer Academic Publisher. ISBN-10: 1-4020-5815-2, ISBN-13: 978-1-4020-5815-8.

Zdzisław Pawlak. 1982. Rough sets. *International Journal of Information and Computer Sciences*, 11(5):341–356.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133. Cambridge University Press.

Sylvain Schmitz and Joseph Le Roux. 2008. Feature Unification in TAG Derivation Trees. In Claire Gardent and Anoop Sarkar, editors, *TAG+9*, pages pages 141–148, Tübingen, Allemagne. 12 pages, 4 figures.

Andrzej Skowron, Jan Komorowski, Zdzislaw Pawlak, and Lech Polkowski, 2002. *Rough sets perspective on data and knowledge*, pages 134–149. Oxford University Press, Inc., New York, NY, USA.

R. Unnikrishnan, C. Pantofaru, and M. Hebert. 2007. Toward objective evaluation of image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):929–944, June.

Marc Verhagen, Robert Gaizauskas, Franck Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 - 15: TempEval Temporal Relation Identification. In *Proceedings of SemEval workshop at ACL 2007*, Prague, Czech Republic, June. ACLP.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pages 45–52, Columbia, Maryland, USA. ACL.

Anne Vilnat, Patrick Paroubek, Eric Villemonte de la Clergerie, Gil Francopoulo, and Marie-Laure Guénot. 2010. Passage syntactic representation: a minimal common ground for evaluation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh LREC*, Valletta, Malta, May. European Language Resources Association (ELRA).

Rebecca Watson, John Carroll, and Ted Briscoe. 2005. Efficient extraction of grammatical relations. In *Proceedings of the Ninth International Workshop on Parsing Technologies (IWPT)*, pages 160–170, Vancouver, October. Association for Computational Linguistics.