Detecting measurement expressions using NooJ

Božo Bekavac
Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
bbekavac@ffzg.hr

Željko Agić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
zagic@ffzg.hr

Krešimir Šojat
Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
ksojat@ffzg.hr

Marko Tadić
Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
mtadic@ffzg.hr

ABSTRACT. *We present a NooJ module implementing a general method for detection and classification of measurement expressions in English and Croatian newspaper texts using local regular grammars. Expressions involving the most frequently used units of measurement are covered for both languages. Insight on module design is provided along with its evaluation. Overall accuracy of the module reaches above 96 and 98 percent for Croatian and English, respectively. Issues regarding normalization of detected measurement expressions are also discussed.*

1 Introduction

Previous successful experiments in developing and evaluating a system for named entity detection and classification in contemporary Croatian texts (cf. Bekavac and Tadić 2007b; Bekavac et al. 2009) using local regular grammars as an underlying formalism have indicated that certain phenomena in natural languages, regardless of their overall complexity, could be successfully modeled in the framework of regular languages. In the experiment presented here, we argue that language expressions involving (units of) measurement in both Croatian and English texts exhibit similar properties. Excluding mere observations and prior linguistic experience, we base this argument on a certain property that is shared between named entities and measure expressions in written language. Namely, both named entities and measure expressions are considered, from this specific perspective, as connection points in which the language system references the physical world. Having previously shown how named entities in Croatian – a highly inflectional and relatively free word-order language – can be efficiently handled by using exclusively local regular grammars and inflectional lexica in INTEX (Silberztein 1999), with an average detection and classification score of over 90 percent on newspaper text, we set off to show how even better scores can be achieved for measurement expressions within the same framework. We based our expectations on overall success of the experiment solely on an intuitive claim that named entities, as defined by the MUC-7 named entity task specification (Chinchor 1997), are a more complex language subsystem when compared with measurement expressions.

The remainder of the paper is structured as follows. First we present the module design and implementation, followed by a brief description of required language resources, namely Croatian and English lexicons and newspaper corpora on which the experiment is conducted. Further on we describe the experiment setup and present the obtained results. We conclude with future research perspectives involving detection and classification of measurement expressions within the presented paradigm.

2 Module design

Before setting off to describe the module itself, we must define the problem of measurement expression detection and classification in a more elaborate manner. Namely, we firstly define what measurement expressions are and how we approach them from a perspective of processing Croatian and English newspaper texts.

Measurement expressions are natural language constructs involving units of measurement. For example, the expression *three hundred miles per hour* is a valid measurement expression in English, denoting velocity of a certain moving object, while the expression *od 7 do 13 stupnjeva Celzijevih* (en. *from 7 to 13 degrees Celsius*) denotes a temperature range in Croatian. In this specific experiment, we detect seven

different types of measurement expressions, namely those involving length, area (surface), volume, weight or mass, pressure, speed (velocity) and temperature. Within these categories, we detect both simple expressions and somewhat more complex expressions, involving ranges of quantities and their language modifiers. For example, we consider *170 square meters* to be a simple length expression and *160-170 square meters* as a more complex one. An expression utilizing modifiers might appear in text as e.g. *from 160 to approximately 170 square meters* and is considered here as the most complex expression form, specifically in terms of normalization. Even though units of measurement are extremely well-elaborated within and beyond our seven simple categories by using the International System of Units (the SI system, cf. Thompson and Taylor 2008), our classification is oversimplified with a purpose. Namely, this experiment is highly task-oriented (or text-oriented), as we do not explicitly seek to detect and classify every measure expression using units of measurement defined by the SI system of units, but rather to achieve high coverage of these expressions in Croatian and English newspaper texts in terms of F1-scores, i.e. maximizing precision and recall on the most frequent unit categories. By targeting common newspaper texts in both languages, in this stage of development, our system will not detect e.g. megawatts and kilojoules, as it will favor e.g. tons and degrees Fahrenheit. This compromise, namely our focus on newspaper domain, is governed by our overall goal of illustrating general principles for measurement detection module design, which can afterwards be easily expanded to cover other units from the SI system and implicitly a broader range of text domains.

The module is developed in NooJ (Silberztein 2003, 2004, 2005) as a sequence of cascaded local regular grammars. Both sets of grammars (the Croatian and English sub-module) utilize NooJ regular expression syntax and lexical resources available for both languages. English lexical resources are distributed within the NooJ package, while their Croatian counterparts are described in (Bekavac et al. 2007a). Each of the grammars consists of four linked sections:

(1) the section detecting numbers and other numerical expressions,

(2) the section detecting metric units of measurement for the given unit class,

(3) the section detecting other, i.e. more frequently used non-metric units of measurement for the given class and

(4) the section that handles frequently used expressions not belonging to any of the previous sections for a given unit class.

This type of modular design was chosen to enable grammar reusability, given that implicit inheritance relationships exist between various units of measurement. As an illustration, expressions denoting surface or area can clearly be viewed from an object-oriented design point of view as adding properties to expressions denoting length. For example, expression *five square meters*, which denotes surface, clearly contains an expression denoting length (*five meters*) and a modifier (*square*) that

adds properties to the given expression, thus making it denote surface. This design principle propagates throughout our local grammars and is illustrated by figure 1.

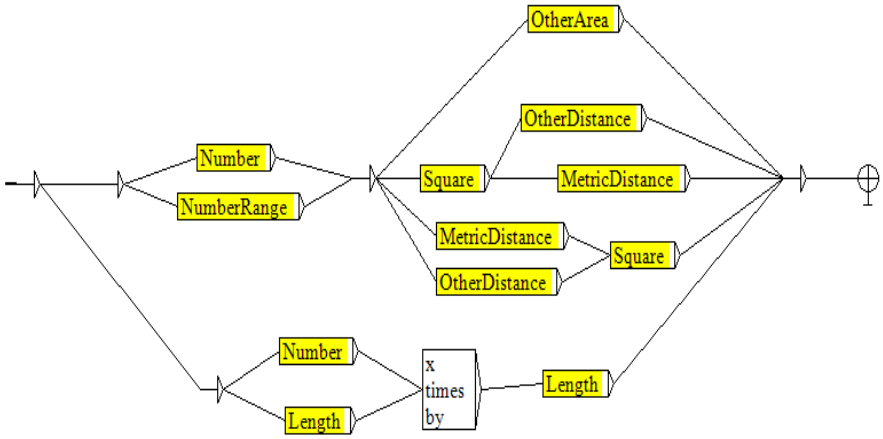

Figure 1. *Local regular grammar for detecting area (or surface) expressions in English*

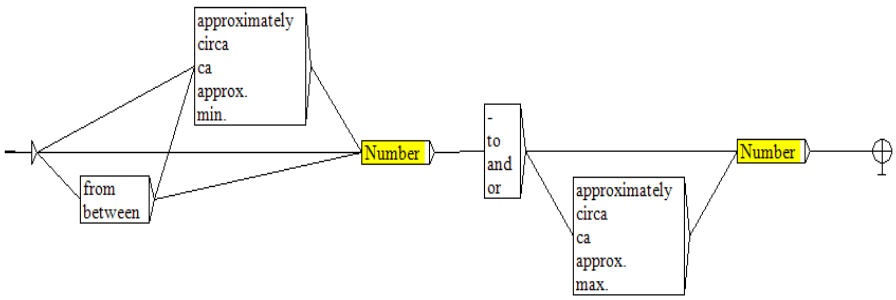

Figure 2.*The NumberRange local grammar for English*

In the figure, the before-mentioned grammar sections are clearly indicated. Local grammars *Number* and *NumberRange* detect numeric expressions and they are fol-

lowed by a set of rules utilizing length-detecting grammars and modifiers contained within the grammar *Square*. The bottom-most branch of the grammar handles certain exceptions. For example, an English language expression such as *100 by 100 miles* is an area expression, even though it is assembled by using two length expressions and a certain multiplication operator. Other local grammars follow exactly the same underlying design principles, both in Croatian and English sub-module.

In the current implementation of the module, only detection and classification of expressions is accounted for, excluding the normalization step. As defined here, normalizing measurement expressions can be approached from two different viewpoints:

(1) normalizing numerical parts of measurement expressions and

(2) normalizing the units of measurement with regards to a certain unit norm.

As an illustration, consider a simple expression in English, e.g. the expression *eleven miles*. From the first viewpoint, keeping in mind a certain information extraction perspective, we should normalize *eleven* as *11*. The second viewpoint requires converting various units of measurement to a certain standard, e.g. the SI system. Therefore, *miles* should be converted to *meters* (*m*). Note that the second viewpoint places an additional constraint on the first one, as *eleven miles* obviously does not equal *11 m*. Miles should thus be converted to meters and this conversion is inherently ambiguous, as the choice of statute miles, metric miles and nautical miles (among others) can only be made in terms of text subject or domain. Moreover, arithmetic operations such as division and multiplication, required to normalize numerical expressions, cannot be performed within the NooJ environment. Furthermore, an even more complex discussion emerges when considering normalization of complex measurement expressions involving number ranges. Namely, how to approach normalization of expressions such as *more than three square miles* or *up to 100 kilometers per hour*? Given all these open and difficult issues, we chose to omit normalization from the scope of the current implementation.

Measurement expression normalization issues and other future work plans are also additionally discussed in the closing section of the paper. In the following section, we proceed with a description of experiment setup and a discussion of obtained results.

3 Experiment and results

The experiment was conducted using the Southeast European Times newspaper corpus (cf. Bekavac et al. 2008 and http://www.setimes.com/). For each of the languages, Croatian and English, the test sample consisted of approximately 1745 articles or approximately 330.000 tokens. No linguistic pre- or post-processing other than the one provided by Croatian and English lexical resources for NooJ was conducted or used throughout the experiment, indicating low resource requirements of

the module. The module was run on this corpus and evaluated manually. The results are presented in table 1.

Before discussing the results in more detail, it should be noted that the absence of gold standard corpora for Croatian and English, annotated with measurement expressions, required us to manually evaluate the module. In turn, manual evaluation made it very difficult (or virtually impossible) to calculate recall on such a large collection given the time and other constraints. Therefore, results are here presented in terms of overall system accuracy or precision on different measurement classes and overall, i.e. average precision scores for the two modules, rather than using the harmonic mean constraining precision and recall, i.e. the F1-measure.

Unit	English	Croatian
Length	90.62	100.0
Area	97.37	100.0
Volume	100.0	100.0
Mass	100.0	94.12
Pressure	100.0	100.0
Speed	100.0	100.0
Temperature	100.0	46.15 (80.77)*
Average	98.28	91.48 (96.41)*

Table1. *Module precision*

On both Croatian and English texts, the overall accuracy of the module is rather satisfactory, as are the individual scores on the seven distinct unit categories. It should be noted that scores on detecting and classifying expressions involving temperature in Croatian texts were somewhat unexpectedly degraded by encountering articles reporting about earthquakes. This was due to the recall-raising relaxed rules that incorrectly classified expressions such as *pet stupnjeva* (en. *five degrees*) as denoting temperature within expressions such as *pet stupnjeva po Richteru* (en. *five degrees on the Richter scale*). Bracketed scores indicate the Croatian module performance when correcting these errors in favor of precision. Including earthquake magnitude and similar measures in the local grammar cascade would also resolve the issue. However, as the scope, i.e. overall goal of this experiment was to illustrate general principles of measurement expression detection using local regular grammars in NooJ and to apply an implemented prototype on Croatian and English newspaper texts, we excluded further classes of measurement units from the presented implementation. Our future work plans might include expanding the overall coverage of measurement unit classes, while the presentation we lay out here might serve as a guideline for prospective readers attempting to implement more complex systems for measurement expression detection.

Besides calculating overall accuracy of the system, we investigated some general properties of errors for both languages. In English texts, it was interesting to note how expressions such as *50 meter butterfly* or *7.65 mm pistol* were classified as length expressions. Classifications of this type are linguistically sound, being that expressions from the examples are indeed valid measurement expressions, taking part in assembling other higher level expressions. However, from a point of view of e.g. specific information retrieval and/or extraction tasks, it could be fatherly argued as to whether detecting such constructs as measurement expressions is indeed useful with regards to overall goals of the larger scale natural language processing system. Indicative errors include expressions such as *additional one hundred meters every day* (speed, not length) or *a person can have one foot in the sea* (not length). Errors on Croatian text exhibited similar properties when compared with the English module, including the before-mentioned issue regarding misinterpretations of earthquake magnitude expressions.

4 Conclusions and future work

We presented a NooJ module for detection and classification of measurement expressions in Croatian and English newspaper texts. Its overall accuracy peaked at around 96 and 98 percent, respectively. Further work on the module might include approaches to internal (NooJ) or external (cf. Bekavac et al. 2009) normalization of measurement expressions for specific IE/IR tasks. External normalization within the referenced paradigm would include designing a measurement expression detection module in INTEX/NooJ, exporting it to a grammar file using one of these development environments and developing additional features, such as expression and unit normalization in one of the programming languages supported by our engine presented in (Bekavac et al. 2009). Such an approach would require fine-tuning the engine to fully support measurement expression detection, as its current version only supports named entity recognition and classification in Croatian newspaper texts.

Other research paths might also be undertaken in order to additionally improve the prototype presented here. Other classes of measurement units – earthquake-related units, units denoting power and energy, etc. – might be included by implementing additional cascades of local regular grammars, thus creating a system able to detect measurement expressions beyond the scope of newspaper texts.

Modules for languages other than Croatian and English might also be implemented using the same design principles and available language resources within the INTEX/NooJ development environment.

5 Acknowledgement

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1002, 130-1300646-1776 and 130-1300646-0645.

References

Bekavac Božo, Vučković Kristina, Tadić Marko. 2007. Croatian Resources for NooJ. 2007 NooJ Conference Book of Abstracts, Xavier Blanco Escoda, Max Silberztein (eds). Barcelona, 2007.

Bekavac Božo, Tadić Marko. 2007. Implementation of Croatian NERC system. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, Special Theme: Information Extraction and Enabling Technologies. Piskorski Jakub, Tanev Hristo, Pouliquen Bruno, Steinberger Ralf (eds). Prague, Association for Computational Linguistics, 2007, pp. 11-18.

Bekavac Božo, Seljan Sanja, Simeon Ivana. 2008. Corpus-Based Comparison of Contemporary Croatian, Serbian and Bosnian. Proceedings of FASSBL6, Tadić Marko, Dimitrova-Vulchanova Mila, Koeva Svetla (eds). Zagreb, Croatian Language Technologies Society, 2008. pp. 33-39.

Bekavac Božo, Agić Željko, Tadić Marko. 2009. Interacting Croatian NERC System and INTEX/NooJ Environment. Proceedings of the 2008 International NooJ Conference. Silberztein Max, Váradi Tamás (eds), Cambridge Scholars Publishing, in press.

Chinchor Nancy. 1997. MUC-7 Named Entity Task Specification. Proceedings of the Message Understanding Conference. Available at the website http://www.itl.nist.gov/iaui/894.02/related_projects/muc/ (last accessed 2009-12-21).

Silberztein Max. 1999. Text Indexing with INTEX. Computers and the Humanities 33:3, Kluwer Academic Publishers.

Silberztein Max. 1999. INTEX: a Finite State Transducer toolbox. Theoretical Computer Science 231:1, Elsevier Science.

Silberztein Max. 2003. NooJ Manual. available at the WEB site http://www.nooj4nlp.net (200 pages).

Silberztein Max. 2004. NooJ : an Object-Oriented Approach. In INTEX pour la Linguistique et le Traitement Automatique des Langues, C. Muller, J. Royaute M. Silberztein Eds, Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comte, pp. 359-369.

Silberztein Max. 2005. NooJ's Dictionaries. In the Proceedings of the 2nd Language and Technology Conference, Poznan : 2005.

Silberztein Max. 2006. NooJ's Linguistic Annotation Engine. In INTEX/NooJ pour le Traitement Automatique des Langues, S. Koeva, D. Maurel, M. Silberztein Eds, Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comte,pp.9-26.

Silberztein Max. 2008. Complex Annotations with NooJ. In the Proceedings of the 2007 International NooJ Conference. Cambridge Scholars Publishing: Newcastle.

Thompson Ambler, Taylor Barry N. 2008. Guide for the Use of the International System of Units. NIST Special Publication 811, 2008 Edition. National Institute of Standards and Technology, US Department of Commerce.