



An Environment for Named Entity Recognition and Translation

**



Wrocław University of Technology

Filip Graliński*
filipg@amu.edu.pl

Krzysztof Jassem*
jassem@amu.edu.pl

Michał Marcińczuk**
michal.marcinczuk@pwr.wroc.pl

Introduction

NERT is a *Named Entity Recognition and Translation* module for *Machine Translation* system. The grammar used to express the recognition and translation rules is based on Spejd formalism (Przepiórkowski 2008).

Role of NERT rules:

1. Recognize and “glue” named entities and other expressions (temporal expressions, legal terms etc.).
2. Provide translation.

NERT rule consists of three parts:

1. Phrase matching pattern - phrase that will be replaced by a translation.
2. Context patterns - optional conditions for the phrase context.
3. Translation rule.

Sample rules

Match : <Art\.> <{NUM}> <ust\.> {NUM} <pkt> <{NUM}> <[Uu]stawy> <z> <dnia> <[0-9]{1,2}> <base~{MonthPL}> <[0-9]{4}>
Action: prepend(Art. \2.\4.\6 of the Act of \10 \11:t \12; sem=document)

Input: Podstawa prawna: Art. 56 ust. 1 pkt 1 Ustawy z dnia 29 lipca 2005
Translation: Legal grounds: Art. 56.1.1 of the Act of 29 July 2005

Match : <1> <base~kwartał> <[0-9]{4}r\.>
Action: prepend(1st quarter of \3:s[-2]; sem=time_period)

Input: 1 kwartale 2008r.
Translation: 1st quarter of 2008

CORP_AFFIX=wiceprezes|akcjonariusz|zarząd|other words used in terms denoting (members of) company bodies
CORP_NAME=<{ProperPL};base!~{CORP_AFFIX}>+
CORP_SUFFIX=S.A.

Match : {CORP_NAME} <{CORP_SUFFIX}>
Left : <{CORP_AFFIX}>
Action: prepend(\1:nom \2; sem=organization)

Input: Wiceprezes Zarządu Banku PKO SA
Translation: Vice-president of the Board of Bank PKO SA

Case study

Text to translate:

Kryterium uznania ww. umowy jako znaczącej jest wielkość kapitałów własnych *Telekomunikacji Polskiej S.A.*

Eng.: *The criteria for acknowledging the agreement mentioned above as significant is the volume of **Telekomunikacja Polska S.A.** ownership capital.*

Machine translation without NERT:

*Criterion of acknowledging the agreement mentioned above as significant a volume of the tangible net worth of the **Polish Telecommunications** is a jsc.*

Problem:

Phrase '*Telekomunikacji Polskiej S.A.*' should be translated as '*Telekomunikacja Polska S.A.*' not as '*Polish Telecommunications* is a *jsc*'.

NERT rule:

UpperPL = [A-ZĄĆĘŁŃÓŚŹŻ]
LowerPL = [a-ąćęłńóśźż]
ProperPL = {UpperPL} {LowerPL} *

Match : <{ProperPL}>+ <S.A.>
Action: prepend(\1:nom \2; sem=organization)

Introduce some definitions to simply the rule.

A sequence of words starting from an upper case.

Take nominative forms of words matched as in the first group.

Set the type of the recognized entity.

Rule matching and applying:

Pattern:	{ProperPL}	{ProperPL}	S.A.
Text:	Telekomunikacji	Polskiej	S.A.
Group and action:	\1:nom		\2
Result:	Telekomunikacja	Polska	S.A.

Table 1. Sample transformation of matched phrase.

Evaluation

Manual by human translators:

TEXT	TYPE	TRANSLATION	Score
wrong	-	-	0
ok	ok	wrong	1
ok	wrong	ok	1
ok	ok	ok	2

Table 2. Scoring the named entity recognition and translation accuracy.

#NE	Max score	Actual score	Precision
3160	6320	5132	81.20%

Table 3. Precision of NERT rules according to the scoring presented in Table 2.

Using METEOR metrics (Banerjee, 2005):

1. Take a bilingual “golden standard” corpus of manually translated texts (S | T), the set of all rules *ALL*, and the set of selected rules *SELECTED*
2. Translate all sentences from the corpus S:
 - 2.1 using rules from *ALL*, obtaining translation *T1*
 - 2.2 using rules from the difference: *ALL - SELECTED*, obtaining translation *T2*
3. Using METEOR metrics:
 - 3.1. Compare *T1* to *T*, obtaining *METEOR(T1)*
 - 3.2. Compare *T2* to *T*, obtaining *METEOR(T2)*
4. If *METEOR(T1) - METEOR(T2) > F1* (positive threshold) then assume *SELECTED* as *useful*
 If *METEOR(T2) - METEOR(T1) > F2* (negative threshold) then assume *SELECTED* as *undesirable*
 Otherwise assume *SELECTED* as *unreliable*

#sentences	Avg. score without NERT	#sentences changed with NERT	Avg. score with NERT
9794	0.577	1461	0.581

Table 4. Precision of NERT rules according to METEOR evaluation.

References (selected)

- Banerjee, Satanjeev and Alon Lavie (2005), METEOR: An Automatic Metric For MT Evaluation With Improved Correlation With Human Judgments. Workshop: On Intrinsic And Extrinsic Evaluation Measures For Machine Translation And/or Summarization
- Przepiórkowski A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. [Partial parsing of Polish] Warszawa: Akademicka Oficyna Wydawnicza EXIT.