PolEval 2020: Information extraction and entity typing from long documents with complex layouts

Filip Graliński

Adam Mickewicz University / Applica.ai

June, 18th

1. INFORMACJE O SPÓŁKACH WCHODZĄCYCH W SKŁAD GRUPY KAPITAŁOWEJ

JEDNOSTKA DOMINUJĄCA – PREZENTACJA SPÓŁKI

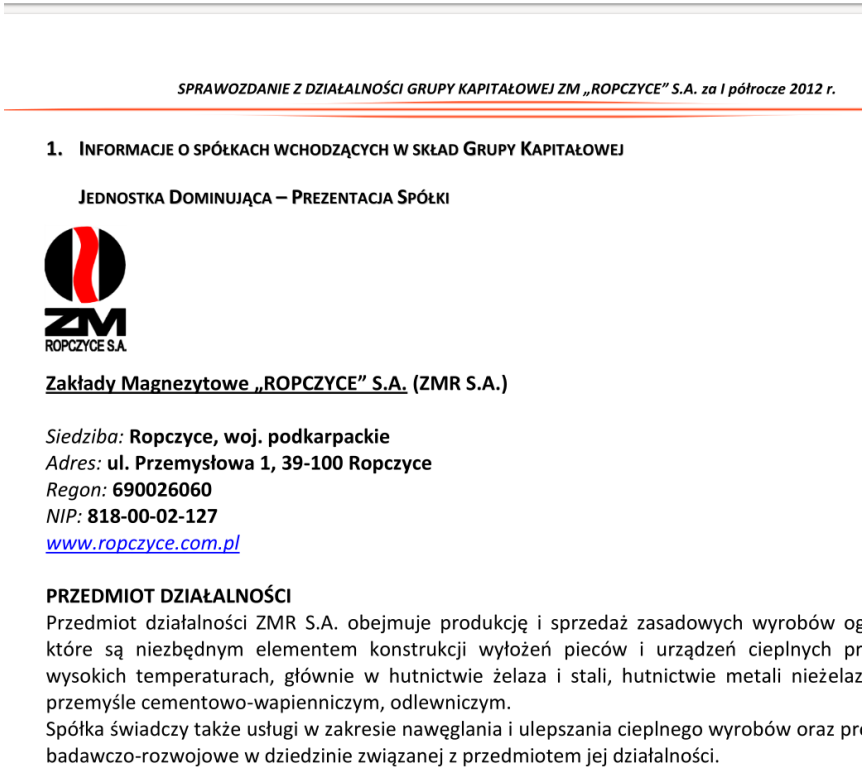
Zakłady Magnezytowe „ROPCZYCE" S.A. (ZMR S.A.)

Siedziba: **Ropczyce, woj. podkarpackie**
Adres: **ul. Przemysłowa 1, 39-100 Ropczyce**
Regon: **690026060**
NIP: **818-00-02-127**
www.ropczyce.com.pl

PRZEDMIOT DZIAŁALNOŚCI

Przedmiot działalności ZMR S.A. obejmuje produkcję i sprzedaż zasadowych wyrobów og
które są niezbędnym elementem konstrukcji wyłożeń pieców i urządzeń cieplnych pr
wysokich temperaturach, głównie w hutnictwie żelaza i stali, hutnictwie metali nieżelaz
przemyśle cementowo-wapienniczym, odlewniczym.
Spółka świadczy także usługi w zakresie nawęglania i ulepszania cieplnego wyrobów oraz pr
badawczo-rozwojowe w dziedzinie związanej z przedmiotem jej działalności.

```
period_from   ?
period_to     ?
postal_code   ?
city          ?
...
```



SPRAWOZDANIE Z DZIAŁALNOŚCI GRUPY KAPITAŁOWEJ ZM „ROPCZYCE" S.A. za I półrocze 2012 r.

1. INFORMACJE O SPÓŁKACH WCHODZĄCYCH W SKŁAD GRUPY KAPITAŁOWEJ

JEDNOSTKA DOMINUJĄCA – PREZENTACJA SPÓŁKI

ZM ROPCZYCE S.A.

Zakłady Magnezytowe „ROPCZYCE" S.A. (ZMR S.A.)

Siedziba: **Ropczyce, woj. podkarpackie**
Adres: **ul. Przemysłowa 1, 39-100 Ropczyce**
Regon: **690026060**
NIP: **818-00-02-127**
www.ropczyce.com.pl

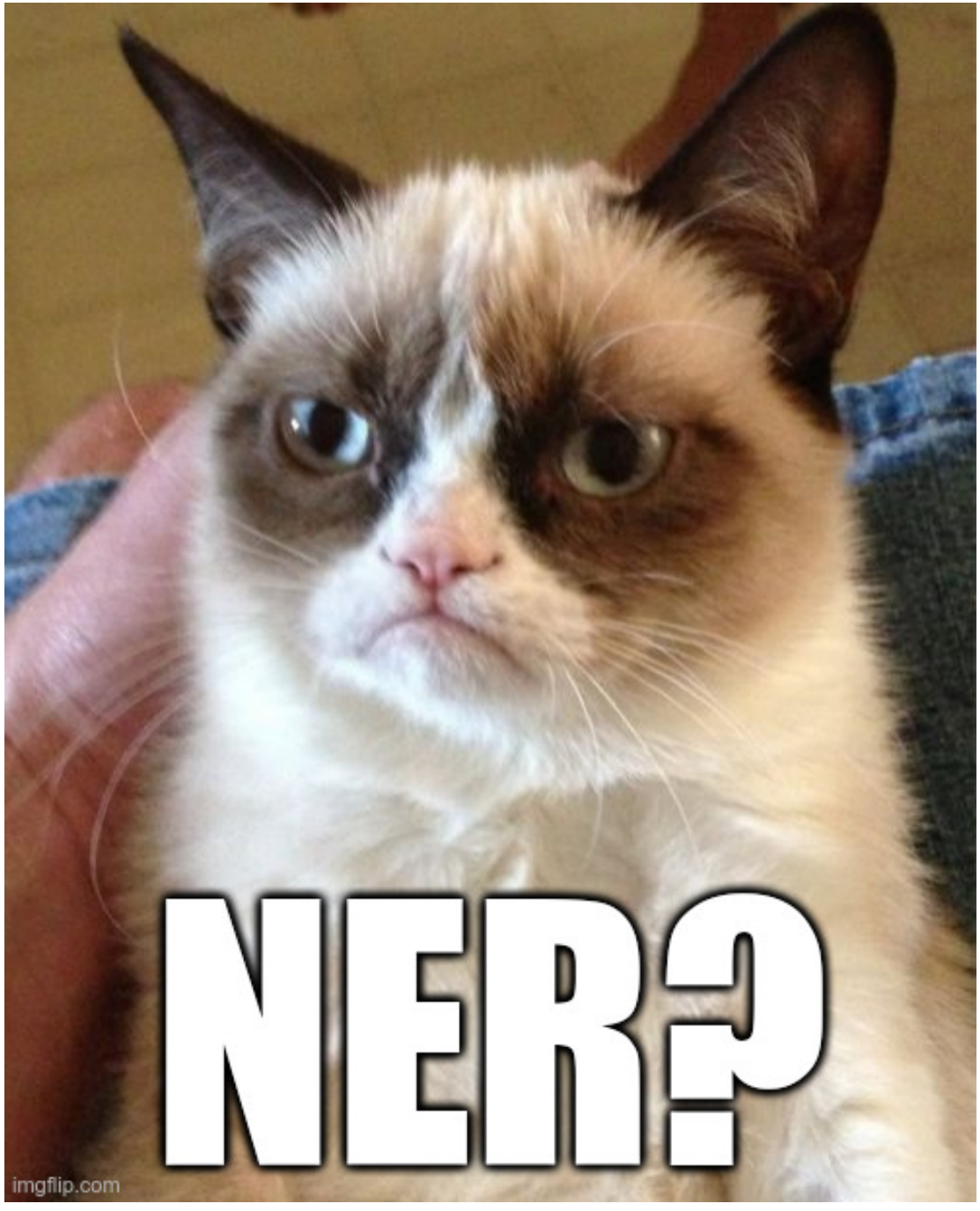
PRZEDMIOT DZIAŁALNOŚCI

Przedmiot działalności ZMR S.A. obejmuje produkcję i sprzedaż zasadowych wyrobów og
które są niezbędnym elementem konstrukcji wyłożeń pieców i urządzeń cieplnych pr
wysokich temperaturach, głównie w hutnictwie żelaza i stali, hutnictwie metali nieżelaz
przemyśle cementowo-wapienniczym, odlewniczym.
Spółka świadczy także usługi w zakresie nawęglania i ulepszania cieplnego wyrobów oraz pr
badawczo-rozwojowe w dziedzinie związanej z przedmiotem jej działalności.

`period_from`	2012-01-01
`period_to`	2012-06-30
`postal_code`	39-100
`city`	Ropczyce

...

`company`, `drawing_date`, `period_from`, `period_to`, `postal_code`, `city`, `street`, `street_no`, `people`


NER?
imgflip.com

... no!

This is an Information Extraction task, not NER*.

- we are interested in the information not where it is
- not just any person, but CEO, etc.

* But of course you could use NER as a part of the pipeline

Evaluation metric

F1-score will be used as the evaluation metric

It's getting complicated for people

```
[('2012-08-30', 'Józef Siwiec', 'Prezes Zarządu'),
 ('2012-08-30', 'Marian Darłak', 'Wiceprezes Zarządu'),
 ('2012-08-30', 'Robert Duszkiewicz', 'Wiceprezes Zarządu')]
```

Hits/failures will be counted for:

- data-point values: `person__name`, `person__position`, `person__signature_date`,
- relations: `person__name__position`, `person__name__signature_date`

Possible approaches

- handcrafted rule, e.g. regexps (as a baseline)

Possible approaches

- handcrafted rule, e.g. regexps (as a baseline)
- standard NER (for general entities) + role classification

Possible approaches

- handcrafted rule, e.g. regexps (as a baseline)
- standard NER (for general entities) + role classification
- specialized NER (but you need to autotag entities)

Possible approaches

- handcrafted rule, e.g. regexps (as a baseline)
- standard NER (for general entities) + role classification
- specialized NER (but you need to autotag entities)
- end-to-end (generative models)

Possible approaches

- handcrafted rule, e.g. regexps (as a baseline)
- standard NER (for general entities) + role classification
- specialized NER (but you need to autotag entities)
- end-to-end (generative models)
- ensembles

Possible approaches

- handcrafted rule, e.g. regexps (as a baseline)
- standard NER (for general entities) + role classification
- specialized NER (but you need to autotag entities)
- end-to-end (generative models)
- ensembles

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Filip Graliński, *LAMBERT: Layout-Aware (Language) Modeling using BERT for information extraction*, `https://arxiv.org/abs/2002.08087`

Anna Wróblewska + students: preparing data set
Dawid Lipiński: scripts for converting into Gonito challenge

Any questions related to the challenge?
Filip Graliński, `filipg@amu.edu.pl`

give it a try? `http://poleval2020.nlp.ipipan.waw.pl/challenge/poleval-financial-reports-pl`

quickly get data (without PDFs):

```
git clone git://gonito.net/poleval-financial-reports-pl
```

hands on!

Assume the following ‘process.py‘ Python script

```
#!/usr/bin/python3

import sys
import re


for line in sys.stdin:
    line = line.replace('\\n', ' ')
    m = re.search(r'(Filip|Anna|Janusz) (\S+)', line)
    if m:
        n = m.group(0)
        n = n.replace(' ', '_')
        print('person__name='+n)
    else:
        print('')
```

Let’s evaluate this simple & stupid solution...x

hands on! cntd.

```
git clone git://gonito.net/poleval-financial-reports-pl
cd poleval-financial-reports-pl
wget https://gonito.net/get/bin/geval
chmod u+x geval
xzcat dev-0/in.tsv.xz | cut -f 3 | ./process > dev-0/out.tsv
geval -t dev-0
```

GEval has many more cool features:
https://gitlab.com/filipg/geval#quick-tour