Extended Named Entitity Recognition in Biomedicine Using Conditionl Random Fields

Monika Žáková

Overview

1. Conditional Random Fields
2. Czech Morphology
3. Extended NER task
4. NER on drug information

Conditional Random Fields

- random field globally conditioned on observation of X
- p(Y|X) constructed without explicitly modelling p(X)
- (X,Y) on G=(V,E) is CRF if Y_v obey Markov property w.r.t. G, i.e.

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$$

X is data sequence, Y is label sequence

w$\sim$v means w is a neighbour of v

CRF II

- if G=(V,E) is a tree

$$p_\theta(y|x) \propto exp(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x)) + \sum_{v \in V,k} \mu_k g_k(v, y|_v, x))$$

- parameter estimation problem

  $\theta = (\lambda_1, ..., \mu_1, ...)$

- $D = (x^{(i)}, y^{(i)})_{i=1}^{N}$ with empirical distribution $\tilde{p}(x, y)$

CRF III

- iterative scaling algorithm maximizing log-likelihood objective function

$$O(\theta) \propto \sum_{x,y} \tilde{p}(x,y) log\, p(y|x)$$

- uses observation-dependent normalization
- slow convergence

Properties of CRF

- gives probabilities of possible labelings given an observation sequence
- in case of fully observable states loss function convex
- features can be on different levels of granularity
- probability of transition not dependent only on current observation
- single exponential model for joint probability of entire sequence of labels given the observation sequence
- avoids label bias problem

Czech Morphology

- morphology for Czech very complex
- high amnount of homonymy e.g. „nehty“ can be 1^{st}, 4^{th}, 5^{th} or 7^{th} case of plural
- standard 15-position tag system
- tags include POS, subPOS, case, gender, number, tense, voice
- in many cases we need to decide on only one tag

Czech Morphology - tools

- tools developed at UFAL UK
- FMorph - dictionary-based tool generating set of possible tags
- taggers - chose one correct tag
- HMM tagger - ability to „guess“ for words not included in dictionary

Named Entitity Recognition

- classical NER: identify names of people, places, companies
- extended NER: identify names of genes, proteins, chemical substances
  1. define classes - classes based on objects not roles, at the same level of granularity, non overlapping
  2. develop annotation guidelines
  3. annotate training set of texts
  4. preprocessing
  5. apply machine learning algorithm

NER in Information about Drugs

- semi structured information about individual drugs
- data set obtained from a database available online on a healthcare portal
- aim: convert most important information into structured form
- e.g. extract information about adverse effects, interactions with other drugs

Defining Classes

- Chemical substance
- Disease e.g. hepatitis
- Disease type e.g. liver disease, liver dysfunction
- Medicine – name of a particular drug
- Medicine type – name of a class of drugs
  e.g. antibiotics
- Measured quantity e.g. blood sugar level
- Symptom e.g. cough, fever
- Treatment e.g. dialysis

Data Set

Initial data set

- Information about 20 drugs
- Drugs of different categories
  - e.g. antibiotics, antiepileptics, antihistamics, etc.
  - 2 documents from each category
- Limited to pills and inhalers
- Basic structure of documents standardized
- Only clinical information considered for experiments

Annotated Named Entities

Class	Number of entities (unique)
Chemical substance	746 (299)
Disease	567 (376)
Disease type	218 (150)
Medicine	141
Medicine type	296
Measured quantity	160
Symptom	797 (516)
Treatment	96
TOTAL	**3022**

Preprocessing

1. Section identification
2. Subsection identification
3. Morphology + Tagger
4. Subsections corrected
5. Morphology + Tagger
6. Processing of numbers and units
7. Add additional features
8. Convert to tab-delimited text

Sections Preprocessing

- Basic sections standard
- Subsections vary with individual documents
- Use of font e.g. bold, italic irregular
- 4 basic types of subheadings identified:
  - Treatment of intoxication
  - Patients with liver disease:
  - Hypertension: Patients with hypertension should be carefully monitored. . . . .
  - Psychiatric disorders: aggressivity confusion, hallucinations,. . .

Available Features

- information about sections and nearest subsection
- number, word, capitalization
- part of speech, case, singular/plural, positive/negative, lemma
- information about class of surrounding tokens

Additional Features

- prefix hypo-, hyper-
- prefix anti-
- sufix -ivum, -ikum
- sufix -émie, -énie, -ida
- sufix -ýza, -úze, -éza, -áza
- sufix -don, -id, -at, -ein, -onin, -oin, -pin, -rin, -gin, -ein, -cin, -zin, -lin, -xim, -min

Additional Features II

- measured quantity: hladina, koncentrace, pocet
- disease type: postizeni, funkce, porucha, onemocneni, infekce
- medicine_type: inhibitor, blokator, antagonista
- disease: akutni, chronicky
- treatment: lecba, terapie, zakrok, vykon, implantace, transplantace, operace
- symptom: obtiz, potiz, stav, zachvat

Learning Using CRF

used tool: CRF++

- open source software
- written in C++ with STL
- marginal probabilities for all candidates
- unigram and bigram features

Feature Templates

```
Input: Data
He          PRP   B-NP
reckons     VBZ   B-VP
the         DT    B-NP << CURRENT TOKEN
current     JJ    I-NP
account     NN    I-NP
```

template	expanded feature
%x[0,0]	the
%x[0,1]	DT
%x[-1,0]	reckons
%x[-2,1]	PRP
%x[0,0]/%x[0,1]	the/DT
ABC%x[0,1]123	ABCthe123

Class Confusions

Disease x Symptom

- e.g. pneumonia, hypertension
- impossible to distinguish from context in some sections
  e.g. unwanted effects

Symptom x Measured quantity

- e.g. high blood pressure x systolic pressure 120
- not enough training data for distinction
- solution: annotate such cases as measured quantity

Chemical substance x Medicine

- too few examples of medicine
- solution: annotate medicine as chemical substance

Results

Class	TP	FP	FN	prec.	recall	F-measure
measured_q.	50	32	21	0.61	0.70	0.65
disease	99	35	126	0.74	0.44	0.55
symptom	201	50	118	0.80	0.63	0.71
treatment	18	1	23	0.95	0.44	0.60
medicine_type	48	1	25	0.98	0.66	0.79
chemical_subst.	166	14	83	0.92	0.67	0.77
disease_type	115	28	37	0.80	0.76	0.78

Discussion

Disease x Symptom

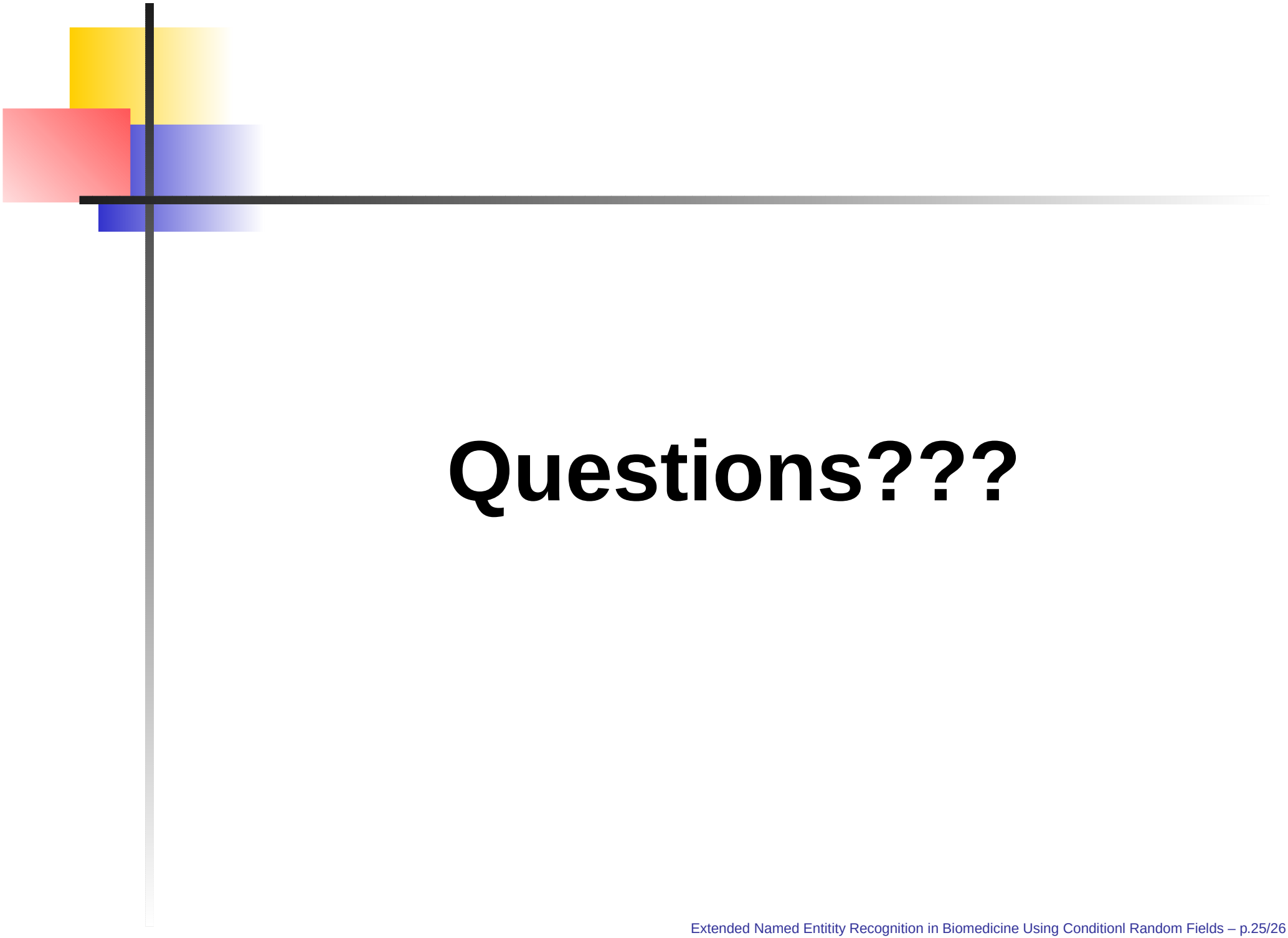
- E.g. pneumonia, hypertension
- Possible solution: for annotation make a decision about class of a term and mark it consistently, not based on context

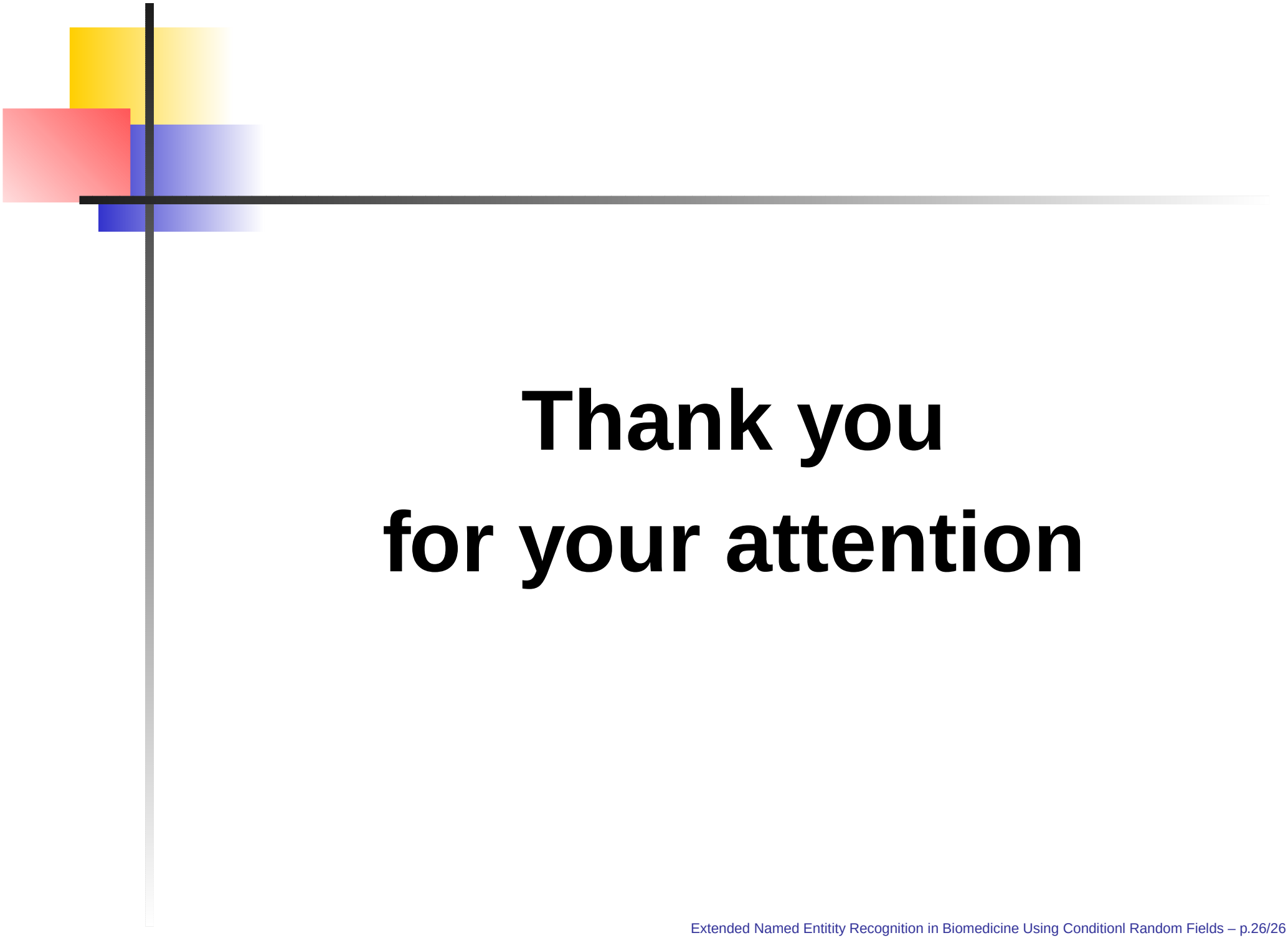
Chemical substance x Measured quantity

- In some cases chemical substance included in measured quantity – in case it appears first as a part of measured quantity, it is marked as such even afterwards
- More examples needed

Ongoing Work

- Updating annotation guidelines for disease and symptom - mark entities as objects not by roles
- Improving lemmatization of words of non-Czech origin
- Working out evaluation scheme
- Extending training data set – use records about drugs of different type with the same distribution as in the database available
- Including more annotators to measure intra-annotator and inter-annotator agreement


Questions???



Thank you for your attention