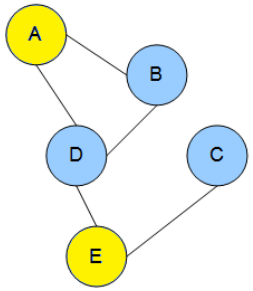
Conditional Random Fields

for Russian NLP tasks

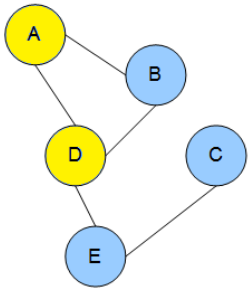
Markov Random Fields (MRF)

Definition. Given an undirected graph $G=(V,E)$, a set of random variables X indexed by V form a Markov random field with respect to G if they satisfy the local Markov properties:

1. Pairwise Markov property: any two non-adjacent variables are conditionally independent given all other variables:

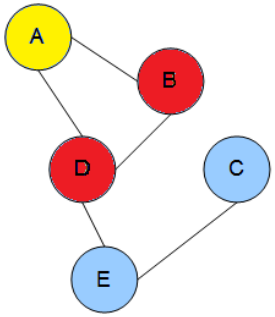


A and *E* are independent given *B, D*, and *C.*



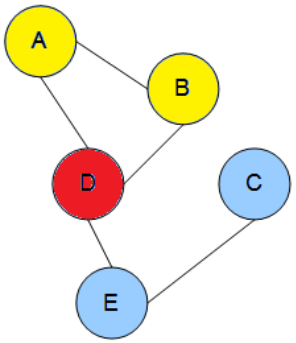
A and *D* are NOT independent given *B,C* and *E.*

2. Local Markov property: a variable is conditionally independent of all other variables given its neighbors:



A is independent on *E* and *C,* given *B* and *D.*

3. Global Markov property: any two subsets of variables are conditionally independent given a separating subset: where every path from a node in A to a node in B passes through S.



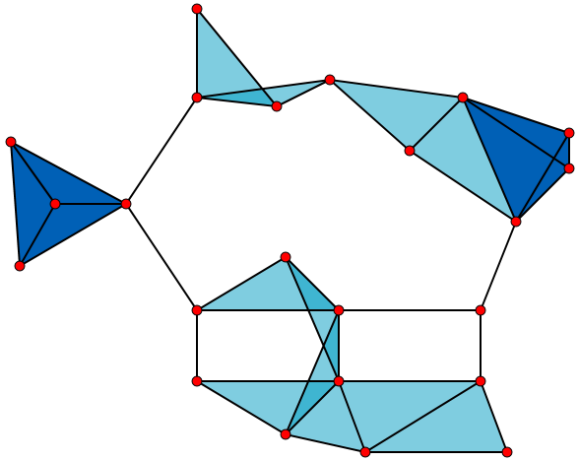
{*A,B*} is independent on {*E,C*} given *D*.

Remark. The above three Markov properties are not equivalent to each other at all. In fact, the Local Markov property is stronger than the Pairwise one, while weaker than the Global one.

Hammersley–Clifford theorem

Definition. Clique in an undirected graph *V* is a subset of its vertices *C* such that every two vertices in the subset are connected by an edge.

Definition. A maximal clique is a clique that cannot be extended by including one more adjacent vertex, that is, a clique which does not exist exclusively within the vertex set of a larger clique.



A graph with 23 1-vertex cliques (its vertices), 42 2-vertex cliques (its edges),
19 3-vertex cliques (the light and dark blue triangles),
and 2 4-vertex cliques (dark blue areas).

Hammersley–Clifford theorem. An undirected graphical model *X* on graph *G* is a Markov Random field if and only if its distribution *P(X)* can be factorized into positive functions defined on cliques that cover all the nodes and edges of *G*. That is,

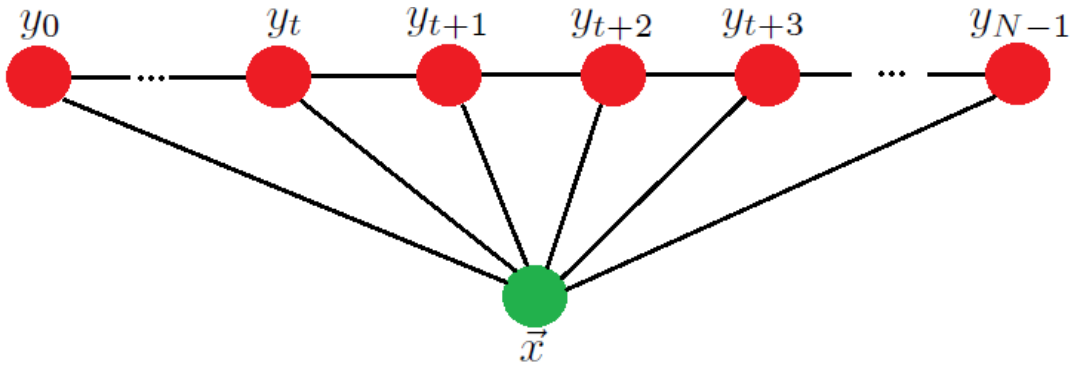
$$P(X) = \frac{1}{Z} \prod_{c \in C_G} \varphi_c(X_c)$$

where C_G is a set of all maximal cliques in *G* and *Z* is a normalization constant.

Remark. φ_c are usually called factors and *Z* is called partition function.

Linear CRF

Consider the following graphical model



The conditional distribution of the model:

$$P(\vec{y} \mid \vec{x}) = \frac{1}{Z(\vec{x})} \prod_{j=1}^{N-1} \Psi_j(y_{j-1}, y_j, \vec{x})$$

If we suppose that

$$\Psi_j(\vec{x}, \vec{y}) = \exp\left\{ \sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}) \right\}$$

Then

$$P(\vec{y} \mid \vec{x}) = \frac{1}{Z(\vec{x})} \exp\left\{ \sum_{j=1}^{N-1} \sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}) \right\}, \text{ where } Z(\vec{x}) = \sum_{\vec{y}} \exp\left\{ \sum_{j=1}^{N-1} \sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}) \right\}$$

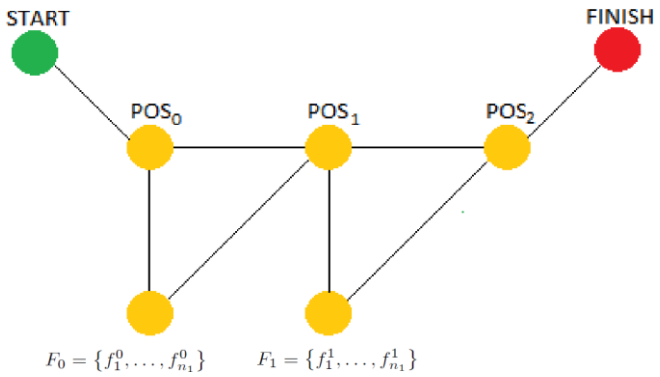
Remark. λ_m are parameters of the model that we should determine during learning.

Results (Morphology and POS-tagging)

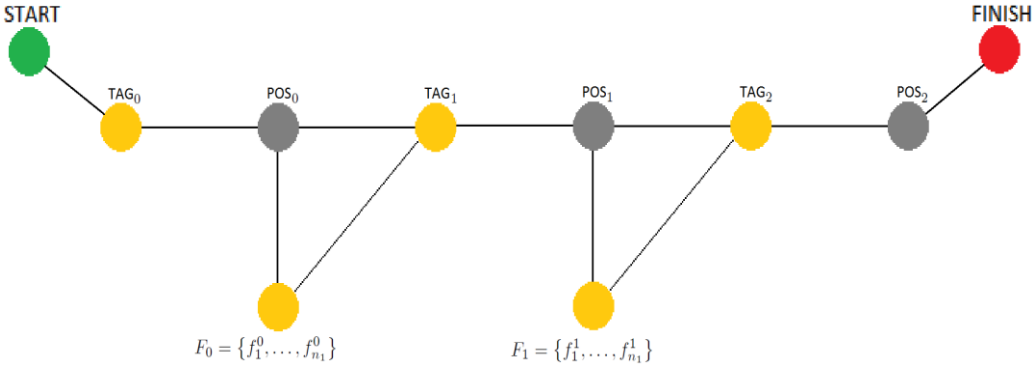
Сорока жила на горе.

СУЩ-ЕД-ЖЕН-ИМ-ОД НУМ-ИМ НУМ-РОД НУМ-ДАТ НУМ-ПР	СУЩ-ЕД-ЖЕН-ИМ-НЕОД ГЛ-НЕСОВ-ИЗЪЯВ-ПРОШ-ЕД-ЖЕН	ПРЕДЛОГ СОЮЗ	СУЩ-ЕД-ЖЕН-ПР-НЕОД СУЩ-ЕД-СР-ИМ-НЕОД СУЩ-ЕД-СР-ВИН-НЕОД СУЩ-ЕД-СР-ПР-НЕОД

Parts of speech (POS) – 11 tags



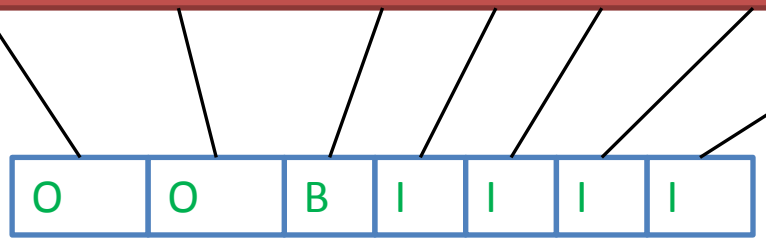
Morphological tags ~ 300-400 tags

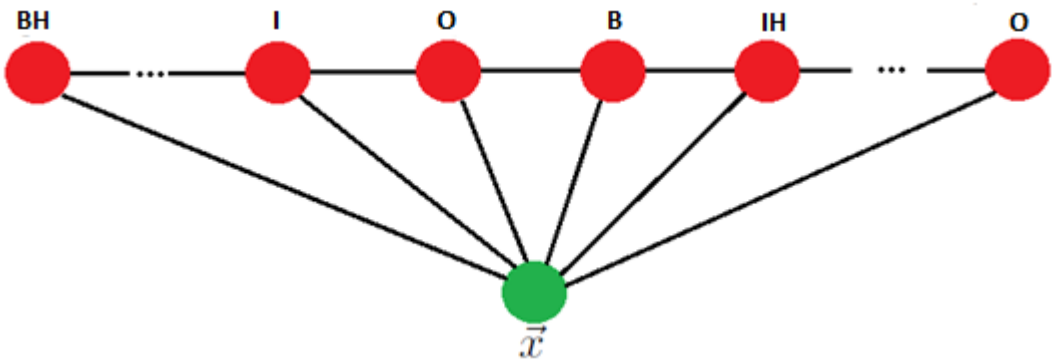


Results(Syntax and Time Expressions detection)

[Президент России Владимир Путин] [голыми руками] поймал [щуку]

Мероприятие закончится [до семи часов вечера завтра]

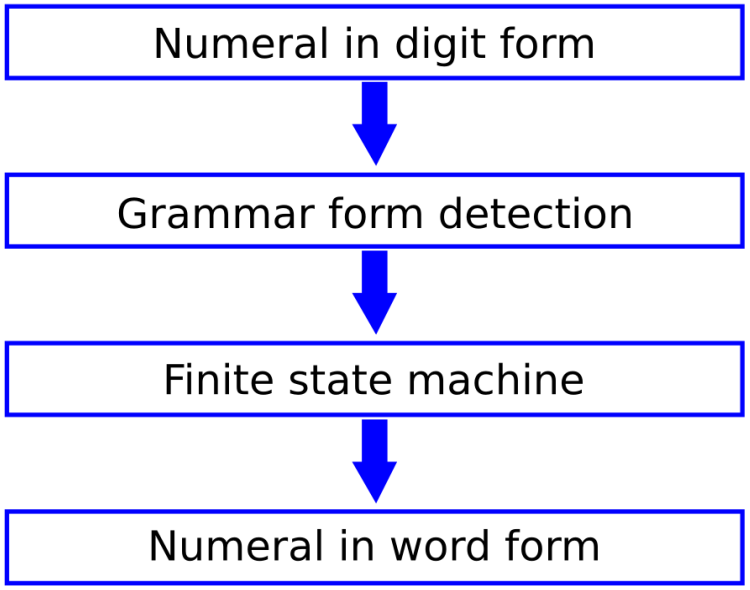




Task of numerals normalization

It is task, which appears in Text-To-Speech System development process.
TTS system should correctly pronounce non-standard phrases, so it should be able to find the word form corresponding to digits and set them to proper grammar form.

Layout of numerals normalization system



Example of normalization

Виктор Ан на олимпиаде 2014 года занял 1 место в забеге на 500 метров.

Виктор Ан на олимпиаде две тысячи четырнадцатого года занял первое место в забеге на пятьсот метров

Results of experiment

Results of quality evaluation on test dataset is Acc = 92.39%

Results of 5-fold cross validation of best configuration is CV=92.21% of accuracy

Result of detection grammar form of numerals averaged by groups of labels

Quality measure	TYPE	CASE	GEN	SNGL	ANIM
P	97.21	91.33	89.77	82.39	87.66
R	97.21	92.93	90.74	85.97	95.05
F_1	97.21	92.10	90.24	84.05	91.11