UNIVERSITY OF TWENTE.

EMPIRICAL CO-OCCURRENCE NETWORKS (CRN) FOR SEQUENCE LABELING

ZHEMIN ZHU, DJOERD HIEMSTRA, PETER APERS, ANDREAS WOMBACHER

DBDBD, 29-11-2013

ROTTERDAM, THE NETHERLANDS

OUTLINE

1. An example of sequence labeling
2. Co-occurrence rate factorization
3. CRN versus CRF
4. Experiments
5. Concluding remarks

APPLICATIONS OF SEQUENCE LABELING

Wide applications:

1. Information extraction
2. Natural language processing
3. Computer vision
4. Bioinformatics
5. …

AN EXAMPLE OF SEQUENCE LABELING
NAMED ENTITY RECOGNITION

I am from Enschede and a member of Dutch Research School for Information and Knowledge Systems.

Potential tags:

- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

This colorful annotation was automatically generated by **Stanford Named Entity Tagger:**http://nlp.stanford.edu:8080/ner/process

AN EXAMPLE OF SEQUENCE LABELING
NAMED ENTITY RECOGNITION

I am from Enschede and a member of Dutch Research School for Information and Knowledge Systems.

Potential tags:

- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

Importance:

1. NER is the first step to extract structured information from unstructured free texts.
2. Named entities are minimum semantic units in many applications.

AN EXAMPLE OF SEQUENCE LABELING
NAMED ENTITY RECOGNITION

I am from Enschede and a member of Dutch Research School for Information and Knowledge Systems.

Potential tags:

- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

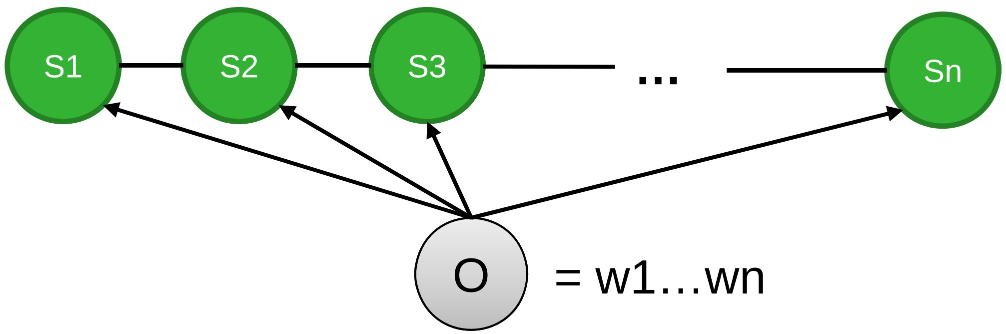
How

Two intuitions:

1. Some words are more likely to be in NER. (Observation Evidence)
2. A word is more likely to be NER if its adjacent words are NER. (Dependence Relation)

AN EXAMPLE OF SEQUENCE LABELING
NAMED ENTITY RECOGNITION

Formalizing the tagging task by probabilistic graphical models:



$$\texttt{argmax}_{s_1,s_2,\ldots,s_n}\, p(S_1, S_2, \ldots S_n \mid O) \quad p \text{ is probability mass function (pmf).}$$

$$\texttt{argmax}_{s_1,s_2,\ldots,s_n}\, f(S_1, S_2, \ldots S_n \mid O) \quad f \text{ is probability density function (pdf).}$$

AN EXAMPLE OF SEQUENCE LABELING
NAMED ENTITY RECOGNITION

Joint probability is not reusable, we need to factorize it.

Known: $p(S_1, S_2 \mid W_1 W_2) \quad p(S_2, S_3 \mid W_2 W_3)$

Predict: $p(S_1, S_3 \mid W_1 W_3)$

AN EXAMPLE OF SEQUENCE LABELING
NAMED ENTITY RECOGNITION

Joint probability is not reusable, we need to factorize it.

Known:

$$p(S_1, S_2 \mid W_1 W_2) = p(S_1 \mid W_1) p(S_2 \mid W_2)$$

$$p(S_2, S_3 \mid W_2 W_3) = p(S_2 \mid W_2) p(S_3 \mid W_3)$$

Predict:

$$p(S_1, S_3 \mid W_1 W_3) = p(S_1 \mid W_1) p(S_3 \mid W_3)$$

CO-OCCURRENCE RATE (CR) FACTORIZATION

Definition of co-occurrence rate (CR):

1. Discrete:

$$\mathrm{CR}(X_1; X_2; ...; X_n) = \frac{p(X_1, X_2, ..., X_n)}{p(X_1)p(X_2)...p(X_n)}$$

2. Continuous:

$$\mathrm{CR}(X_1; X_2; ...; X_n) = \frac{f(X_1, X_2, ..., X_n)}{f(X_1)f(X_2)...f(X_n)}$$

CO-OCCURRENCE RATE (CR) FACTORIZATION

- Partition Theorem

$$\text{CR}(X_1; ...; X_j; X_{j+1}; ...; X_n)$$
$$= \text{CR}(X_1; ...; X_j)\text{CR}(X_{j+1}; ...; X_n)\text{CR}(X_1 ... X_j; X_{j+1} ... X_n)$$

- Cancellation Theorem

If $X \perp Y \mid Z$, then $\text{CR}(X; YZ) = \text{CR}(X; Z)$.

CO-OCCURRENCE RATE (CR) FACTORIZATION

- Factorizing using CR

$$p(S_1, S_2 \dots, S_n \mid O) = \prod_{i=1}^{n} p(S_i \mid O) \prod_{j=1}^{n-1} \mathrm{CR}(S_j ; S_{j+1} \mid O),$$

$$f(S_1, S_2 \dots, S_n \mid O) = \prod_{i=1}^{n} f(S_i \mid O) \prod_{j=1}^{n-1} \mathrm{CR}(S_j ; S_{j+1} \mid O)$$

RELATION BETWEEN CR AND COPULA

- (Bivariate) Copula

$$C_{F_X,F_Y}(u,v) = P(F_X \leq u, F_Y \leq v)$$

F_X and F_Y be the commutative distribution functions (cdf) of X and Y

- Copula density function

$$c_{F_X,F_Y} = \frac{\partial^2}{\partial F_X \partial F_Y} C_{F_X,F_Y}$$

RELATION BETWEEN CR AND COPULA

- Continuous CR is just the Copula density function

Since:

$$f_{X,Y}(x,y) = c_{F_X,F_Y}(F_X(x),F_Y(y)) \begin{vmatrix} \frac{\partial F_X}{\partial X} & \frac{\partial F_X}{\partial Y} \\ \frac{\partial F_Y}{\partial X} & \frac{\partial F_Y}{\partial Y} \end{vmatrix}$$
$$= c_{F_X,F_Y}(F_X(x),F_Y(y)) f_X(x) f_Y(y),$$

Then $\mathrm{CR}_{X;Y}(x,y) = \dfrac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} = c_{F_X,F_Y}(F_X(x),F_Y(y))$

- Discrete CR: Do not know.

RELATION BETWEEN CR AND COPULA

Estimating Copula:

Estimating every marginal distribution, then plug into the joint distribution.

We use a similar idea to estimate CR.

CRN VS. CRF

CRF: $$p(S_1, S_2 \ldots, S_n \mid O) = \frac{1}{Z_O} \prod_{i=1}^{n} \phi_i(S_i, O) \prod_{j=1}^{n-1} \psi_j(S_j, S_{j+1}, O)$$

$$Z_O = \sum_{S_1, S_2, \ldots, S_n} \left[\prod_{i=1}^{n} \phi_i(S_i, O) \prod_{j=1}^{n-1} \psi_j(S_j, S_{j+1}, O)\right]$$

1. Global normalization.
2. Normalizer is very important because it implies constraints.

CRF

CRF: $$p(S_1, S_2 ..., S_n \mid O) = \frac{1}{Z_O} \prod_{i=1}^{n} \phi_i(S_i, O) \prod_{j=1}^{n-1} \psi_j(S_j, S_{j+1}, O)$$

CRN: $$p(S_1, S_2 ..., S_n \mid O) = \prod_{i=1}^{n} p(S_i \mid O) \prod_{j=1}^{n-1} \mathtt{CR}(S_j; S_{j+1} \mid O),$$

Global normalization vs Local normalization

1. ϕ and ψ are not probability distributions and cannot be locally normalized.
2. p is probability distribution, can be locally normalized. CR can be estimated locally using Copula techniques.

EXPERIMENTS

Dataset: Brown corpus for Part-of-speech tagging

Software: CRF++ version 0.57, CRN is implemented by us.

TABLE III: Accuracy On POS Tagging

	Overall	Known	Unknown	Time (Sec.)
CRF++	95.4	96.1	**71.7**	4,571,807
ECRN	**95.6**	**96.9**	70.5	**3.9**

EXPERIMENTS

Dataset: Dutch part of CoNLL-2002 Named Entity Recognition Dataset

Software: CRF++ version 0.57, CRN is implemented by us.

TABLE IV: Accuracy On NER

	Overall	Known	Unknown	Time (Sec.)
CRF++	96.13	98.2	**77.4**	794
ECRN	**96.23**	**98.8**	73.7	**1.3**

CONCLUSION

- We proposed the new Co-occurrence Rate factorization for undirected graphs.
- Local method can be trained much faster and obtain competitive or better results than traditional global methods.

THE END

Thanks you!

This work has been supported by COMMIT/.