BUILDING YOUR WEB NER MODEL VIA SEMISUPERVISED SEQUENCE LABELING

Dr. Chia-Hui Chang
WIDM Lab @ National Central University

Co-work with Chien-Lung Chou, Ya-Yun Huang, Yuan-How Lin

INTRODUCTION

- Named entity recognition (NER) is a fundamental task for many text mining applications
- Labeled training data is expensive and is often limited.
  - CoNLL 2003 shared task provided 14,987 sentences
  - #Microposts2014 workshop provided 2,340 tweets
- Existing Chinese NER models are trained from a small set of news articles
  - Model: SVM, HMM, CRF, or Mixed Model
  - Specific features: POS, Chunking, and Word Segmentation

INTRODUCTION – NER PACKAGE COMPARISON

Table I. NER packages comparison

Category	NER Package	Precision	Recall	F-measure
Chinese Person Name	FundanNLP	0.636	0.688	0.661
	Stanford NER	0.758	0.762	0.760
	Our System	0.936	0.887	0.911
Chinese Biz Org. Name	FundanNLP	0.429	0.081	0.136
	Stanford NER	0.518	0.542	0.530
	Our System	0.825	0.875	0.849
Chinese Location Name	FundanNLP	0.353	0.377	0.365
	Stanford NER	0.215	0.188	0.201
	Our System	0.925	0.777	0.845

RELATED WORK

- Supervised Sequence Labeling
  - HMM, MEMM, CRF
- Distant Learning
  - FreeBase (Relations), Wikipedia Title, FourSquare and Gowalla (POI)
  - English-Chinese discourse level aligned parallel corpus
- Semisupervised learning with unlabeled data
  - Self-Learning, S^3VM (Transductive SVM) for Classification
  - Co-Training / Tri-Training
- Semisupervised learning for Sequence Labeling

SEMISUPERVISED SEQUENCE LABELING

- **Distant Supervision:**
  - Automatic Labeling based on existing known entities to obtain more labeled training data [An et al. 2003]
- **Tri-Training**
  - Making use of unlabeled data via tri-training
- **Sequence Labeling**
  - We use CRF and general features (No Word Segmentation and POS features)

We will discuss three issues:

- How to collect a lot of good quality training data?
- How to apply Tri-Training on large scale data set?
- How to find features for sequence labeling?

Issue 1

How to collect a lot of good quality training data

AUTOMATIC LABELING

- We use automatic labeling and self-testing to solve this issue
- Automatic Labeling
  - Collecting known entities as query keywords
  - Collecting sentences that contain keyword from top 10 query results via search engine / FaceBook / PTT posts
  - Using the all the known entities to label the collected sentences (called full-labeling)

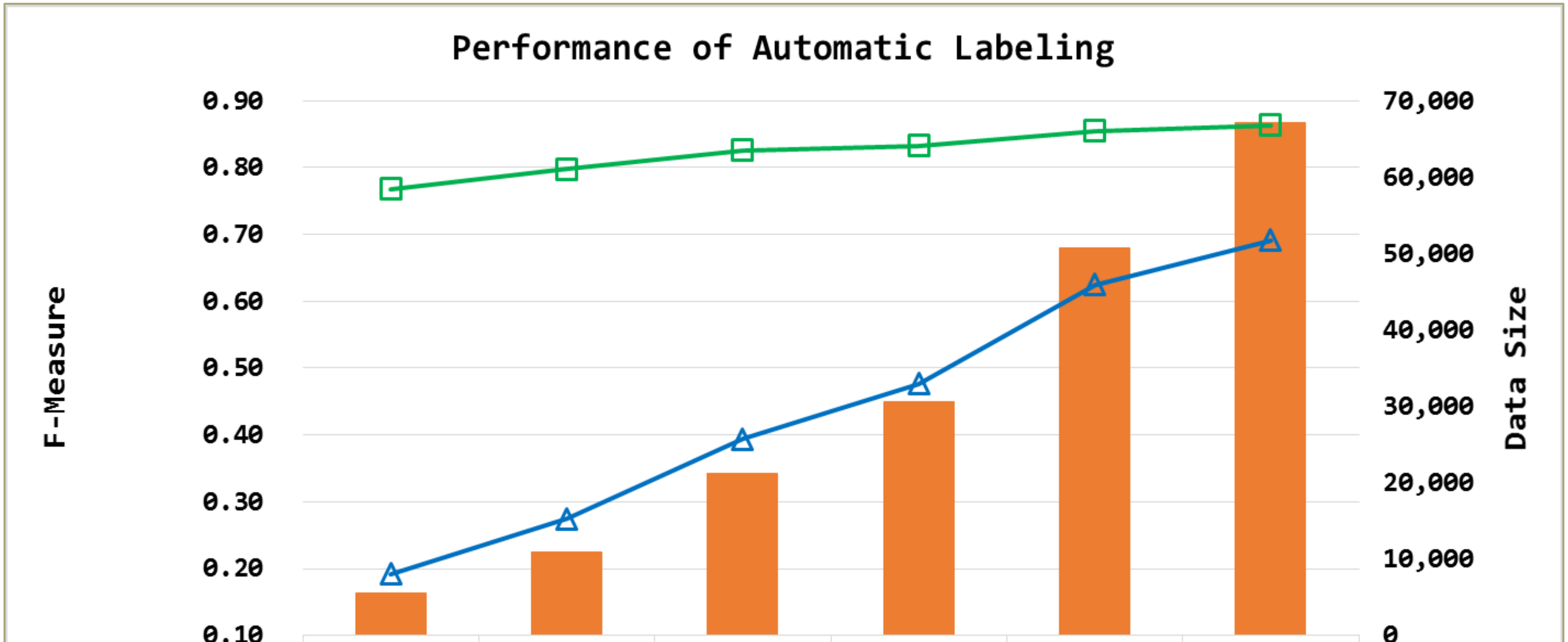
TRAINING/TESTING DATA

	Training Data					
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6
Celebrity	500	1000	2000	3000	5000	7053
Sentences	5548	10928	21267	30653	50738	67104
Words	106,535	208,383	400,111	567,794	913,516	1,188,822

Testing Data

- Collecting news articles from four online news websites
- It contain 11 categories (including politics, finance, sports …) during 2013/01/01 to 2013/03/31
- Total include 8,672 documents, 364,685 sentences, 54,449 person names, and 11,856 distinct person names

AUTOMATIC LABELING PERFORMANCE



	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6
Data Size	5,548	10,928	21,267	30,653	50,738	67,104
Full-Labeling	0.7677	0.7974	0.8254	0.8329	0.8544	0.8636
Uni-Labeling	0.1916	0.2750	0.3939	0.4762	0.6249	0.6916

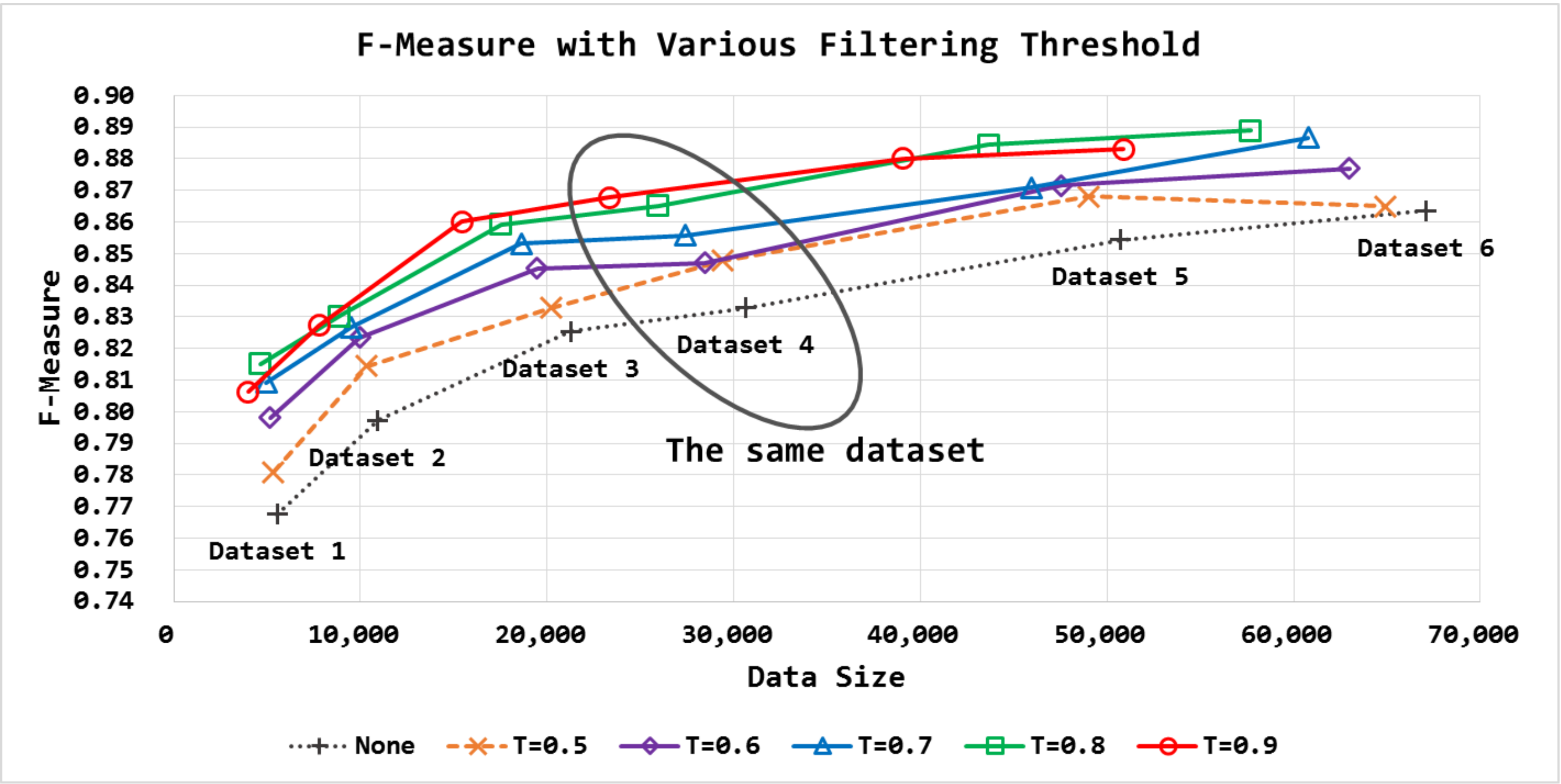
SELF-TESTING

- Remove noise in the automatically labelled training data
  - Using CRF model to test itself (training data) and output the conditional probability
  - Remove sentence with low confidence
  - Threshold = 0.5、0.6、0.7、0.8、0.9

Effects:

- Threshold ↑, |Training Data| ↓ and Data Quality ↑,
- Threshold < 0.8: F-measure ↑
- Threshold > 0.8: F-measure ↓

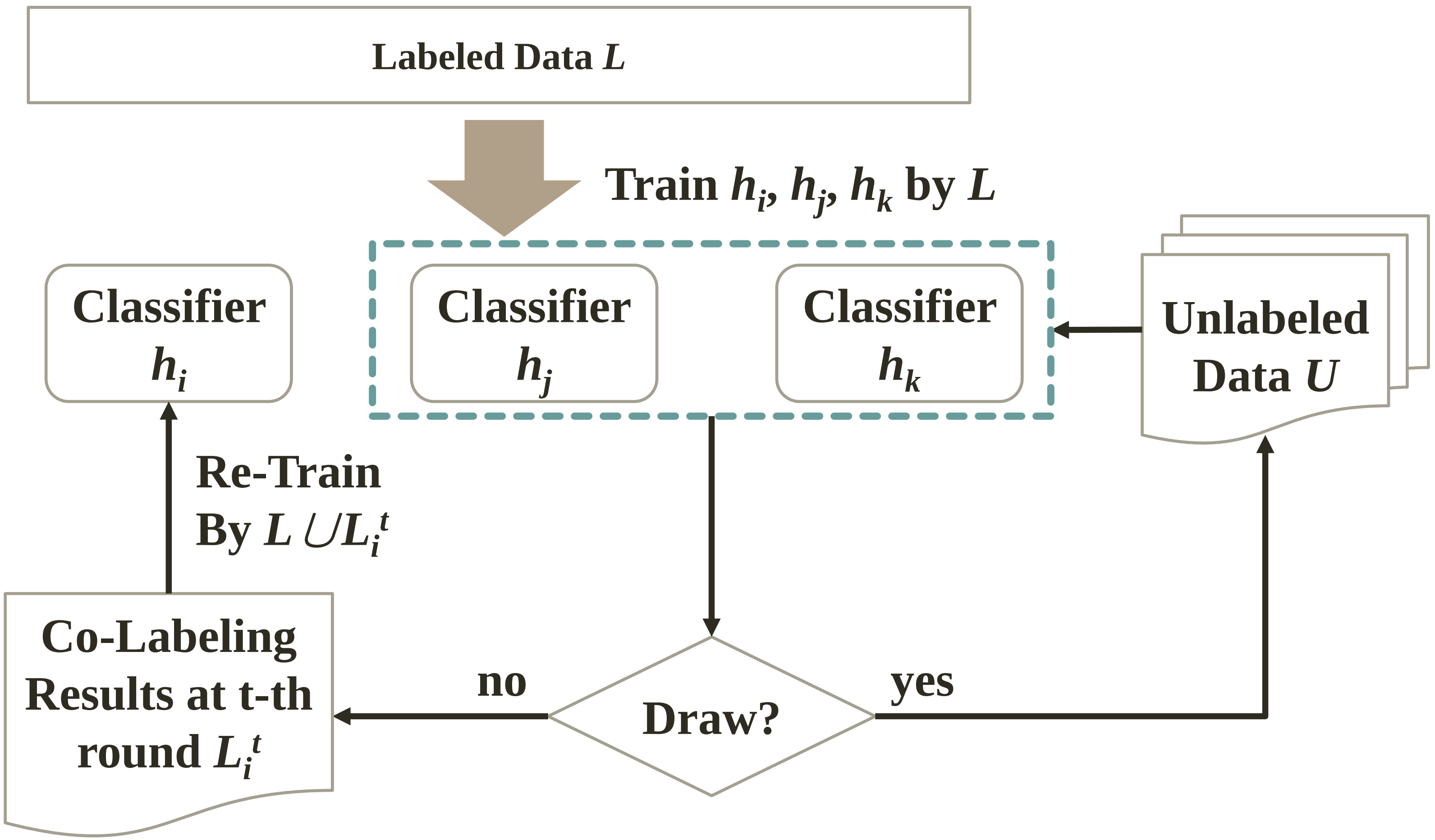
SELF-TESTING PERFORMANCE



Issue 2

How to apply Tri-Training on large scale data set?

TRI-TRAINING



TRI-TRAINING (CONT.)

According to PAC Learning,

- *Learning from noisy examples* proposed by Angluin and Laird in 1988
- To ensure the error rate is reduced through iterations, when training h_i, Eq. (1) must be satisfied,
- $e_i^t \left| L_i^t \right| < e_i^{t-1} \left| L_i^{t-1} \right|$ (1)

$where\, e_i^t$ denotes the error rate of h_i in t-th round on labeled data L

- $e_i^t = \frac{|\{(x,y) \in L,\, h_j^t(x) = h_k^t(x) \neq y\}|}{|\{(x,y) \in L,\, h_j^t(x) = h_k^t(x)\}|}$ (2)
- $L_i^t = \{(x, y): x \in U, y = h_j^t(x) = h_k^t(x)\}$) for model h_i ($i, j, k \in \{1,2,3\}$, $i \neq j \neq k$)

TRI-TRAINING (CONT.)

If $\left|L_i^t\right|$ is too large, Eq. (1) will be violated

We could use Eq. (1) to derivation Eq. (3) to estimate the upper bound u for $\left|L_i^t\right|$

- $u = \left\lceil \frac{e_i^{t-1}\left|L_i^{t-1}\right|}{e_i^t} - 1 \right\rceil$ (3)
- $S_i^t = \begin{cases} Subsample\left(L_i^t, u\right) & violated\ Eq.(1) \\ L_i^t & otherwise \end{cases}$ (4)

$L \cup S_i^t$ is used as training data to update classifier h_i for this iteration.

TRI-TRAINING INITIALIZATION ISSUE

- In order to estimate the size of $\left|L_i^1\right|$, we need to estimate e_i^0, e_i^1, and $\left|L_i^0\right|$ first.

- Zhou et al. assumed a 0.5 error rate for e_i^0, computed e_i^1 by h_j and h_k, and estimated the lower bound for $\left|L_i^0\right|$
  - $\left|L_i^0\right| = \left\lfloor \frac{e_i^1}{e_i^0 - e_i^1} + 1 \right\rfloor = \left\lfloor \frac{e_i^1}{0.5 - e_i^1} + 1 \right\rfloor$ (6)

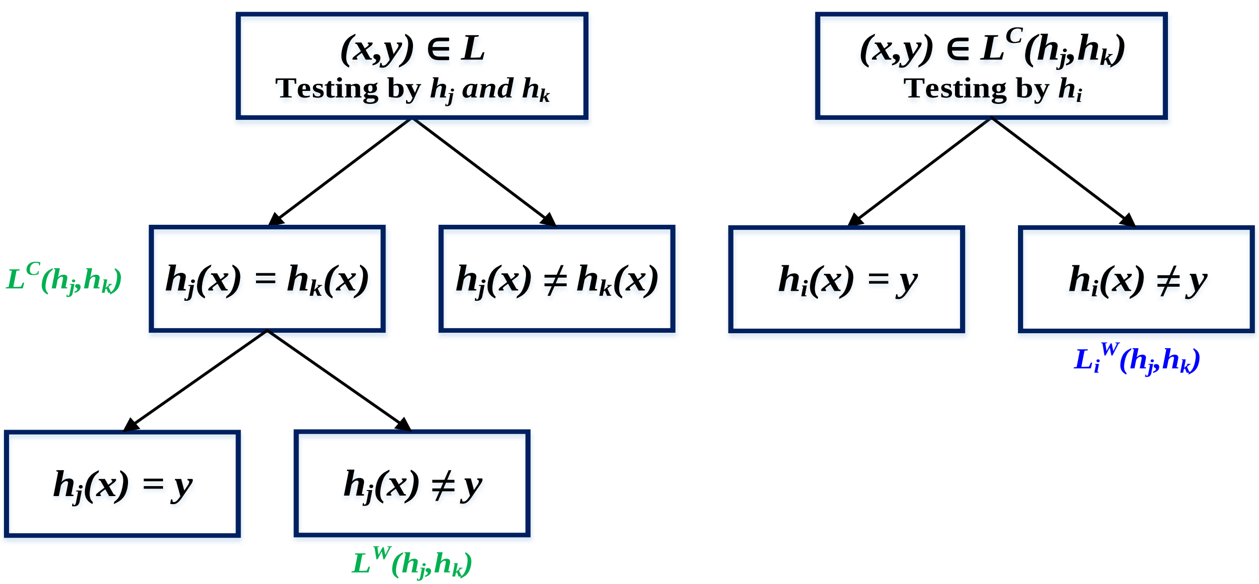
e_i^1	0 ~ 0.4	0.49	0.499	0.4999
$\lvert L_i^0 \rvert$	1 ~ 5	50	500	5000

  - for a larger dataset L, such an initialization $\left|L_i^0\right|$ will have no effect on retraining and will lead to an early stop

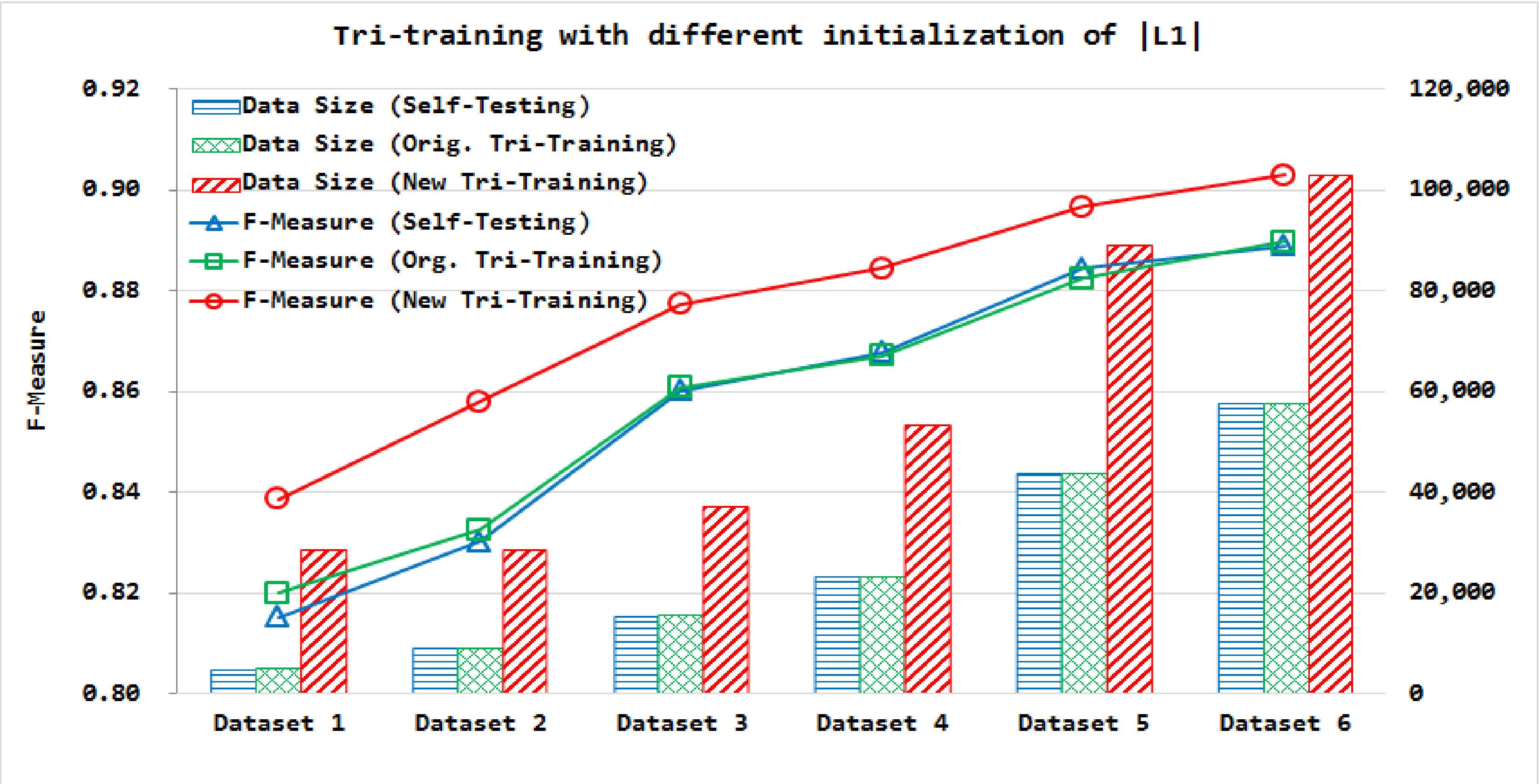
TRI-TRAINING INITIALIZATION ISSUE (CONT.)

In this paper, we propose

- $|L_i^0| = \left\lceil \frac{e_i^0 |L_i^0|}{e_i^1} - 1 \right\rceil = \left\lceil \frac{|L_i^W(h_j,h_k)| * |L^C(h_j,h_k)|}{|L^W(h_j,h_k)|} - 1 \right\rceil$ (10)



TRI-TRAINING PERFORMANCE



Experiments & Applications

OTHER NER TASKS

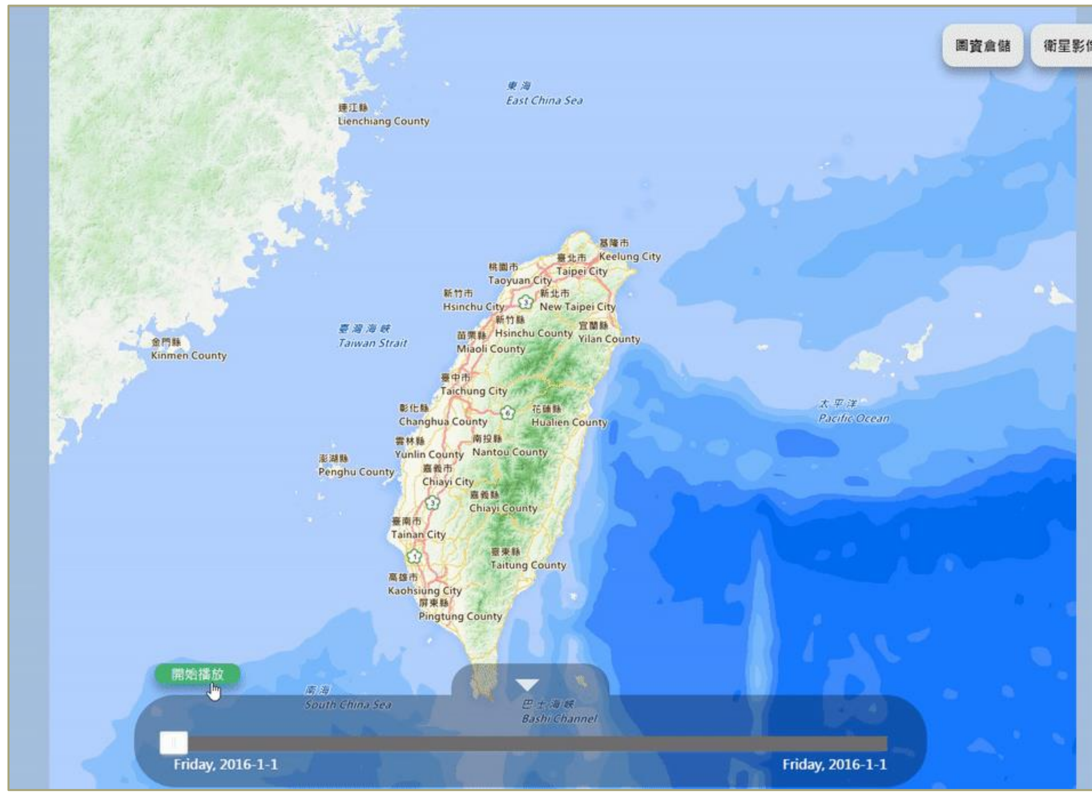
Performance Measure: Partial Match vs. Exact Match

Category	Steps	Precision	Recall	F-measure
Chinese Location	Full-Labeling	0.896	0.769	0.828
	Self-Testing	0.900	0.776	0.833
	Tri-Training	0.925	0.777	0.845
Chinese Biz Org.	Full-Labeling	0.850	0.779	0.813
	Self-Testing	0.808	0.859	0.833
	Tri-Training	0.825	0.875	0.849
English Biz Org.	Full-Labeling	0.781	0.835	0.807
	Self-Testing	0.774	0.868	0.818
	Tri-Training	0.789	0.881	0.832
Japanese Biz Org.	Full-Labeling	0.824	0.730	0.774
	Self-Testing	0.841	0.745	0.789
	Tri-Training	0.845	0.766	0.803

EVENT MONITORING FROM USER-GENERATED CONTENT ON SOCIAL MEDIA

- **FB Event Watch**
  - **Activity name**
  - **Location**
  - **Date/Time**

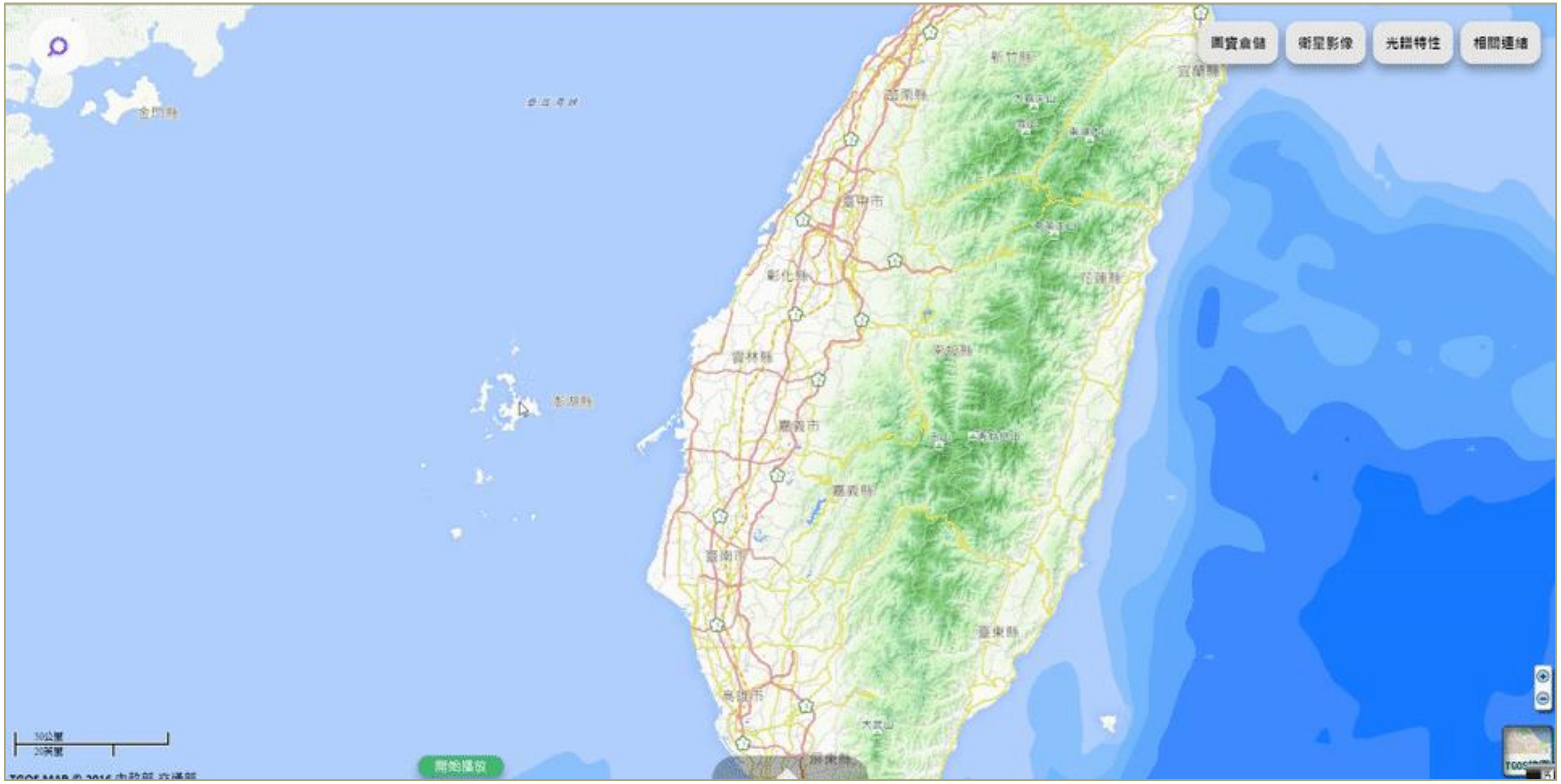
- **Damage Monitoring**
  - **Damage Report**
  - **Location**
  - **Date/Time**



SEARCH FOR POSTS WITH DAMAGE REPORTS?

[Video Demo on YouTube]

Specify a keyword to query the system.



CONCLUSION

- Semi-supervised Sequence Labeling
  - Distant Learning + Tri-Training + Sequence Labeling
  - While such data may contain noise, the benefit with large labeled training data still is more significant than noise it inherits.
- Steps
  1. Seed lists
  2. Text Source: FB/PTT Posts, Search snippets, News articles, etc.
  3. Model Training/Testing
- Release / Sharing of NLP Tools
  - Academic: NER API, Partial package
  - Commercial: Trained NER model, Package for building your own NER Model

Thank you for listening!