Research on conversation thread detection

Yang Wang

CSLT / RIIT

Tsinghua University

wangyang@cslt.riit.tsinghua.edu.cn

目 录

1. Background

2. Motivation

3. Evaluation Metrics

4. Base model

5. Word2vec

6. Classifier

7. Cluster

8. Evaluation Method

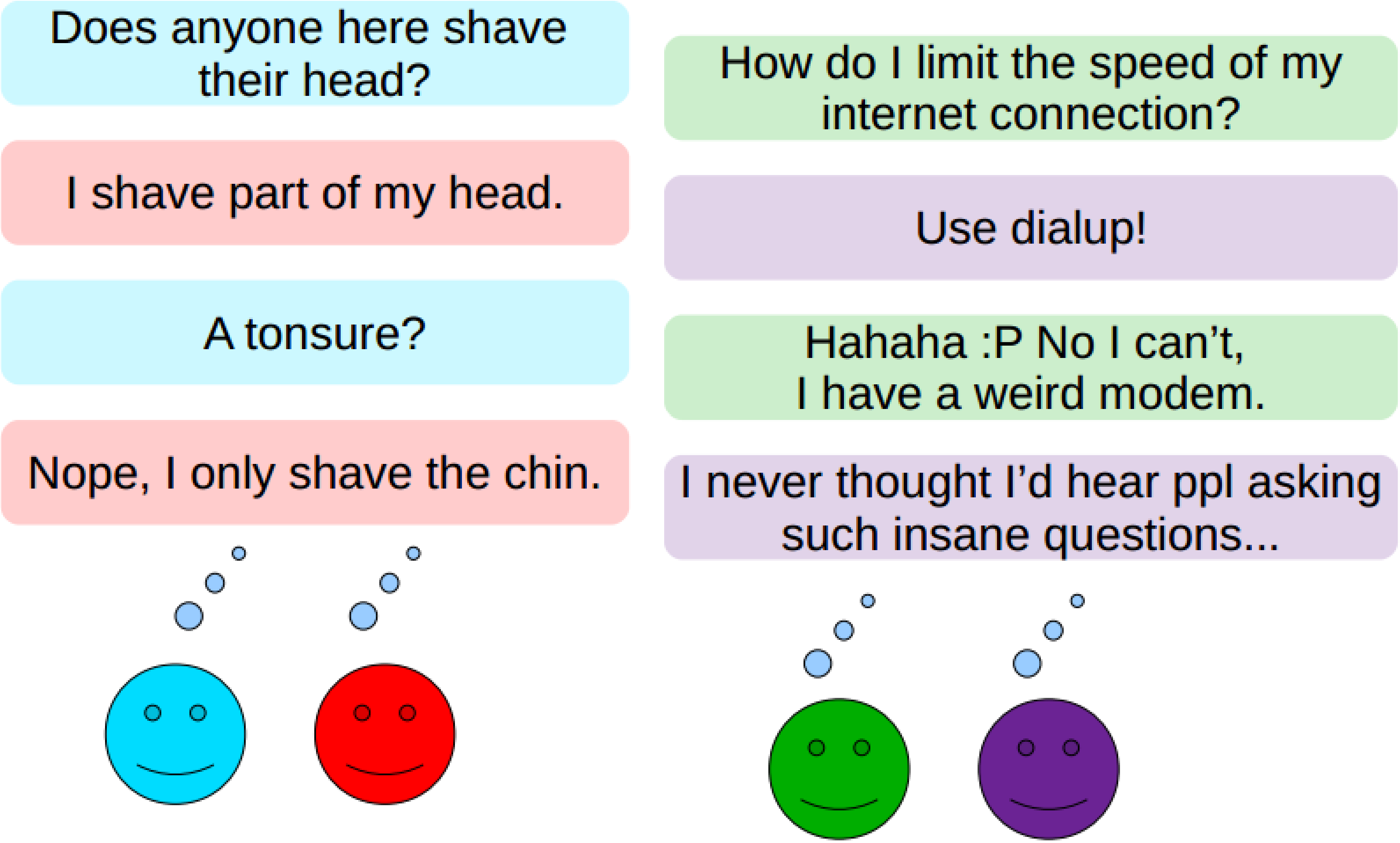
9. Reference

1. Background

- Dynamic text message streams are rapidly growing on the Internet.
- There is a remarkable category of streams containing valuable knowledge.
- There may be more than one thread at the same time, and the text of different threads intersects with each other.

1. Background

Two User Conversation



Does anyone here shave their head?

I shave part of my head.

A tonsure?

Nope, I only shave the chin.

How do I limit the speed of my internet connection?

Use dialup!

Hahaha :P No I can’t,
I have a weird modem.

I never thought I’d hear ppl asking such insane questions...

1. Background

Multi-User Conversation

Does anyone here shave their head?

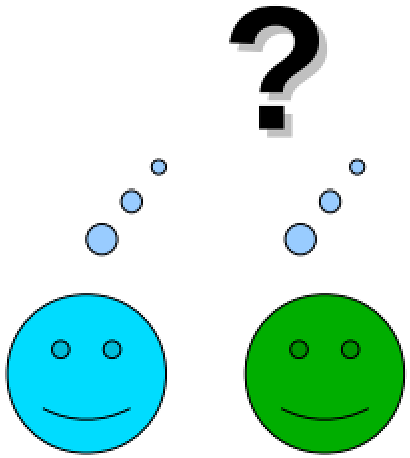
How do I limit the speed of my internet connection?

I shave part of my head.

A tonsure?

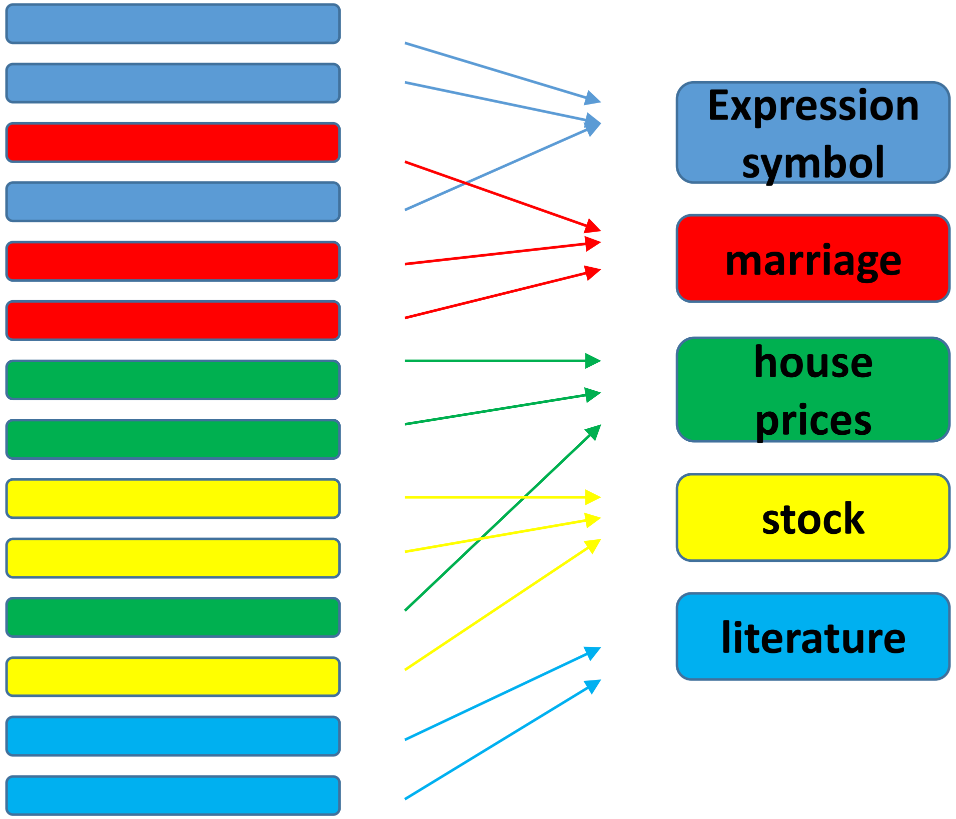
Use dialup!

Nope, I only shave the chin.



- A common situation:
  - Text chat
  - Push-to-talk
  - Cocktail party

2. Motivation



2. Motivation

- A natural discourse task.
  - Humans do it without any training.
- Preprocess for search, summary, QA.
  - Recover information buried in chat logs.
- Online help for users.
  - Highlight utterances of interest.
- Dialogue system with memory.
  - Extract context information.
  - Process data for dialogue system training.

3. Evaluation Metrics

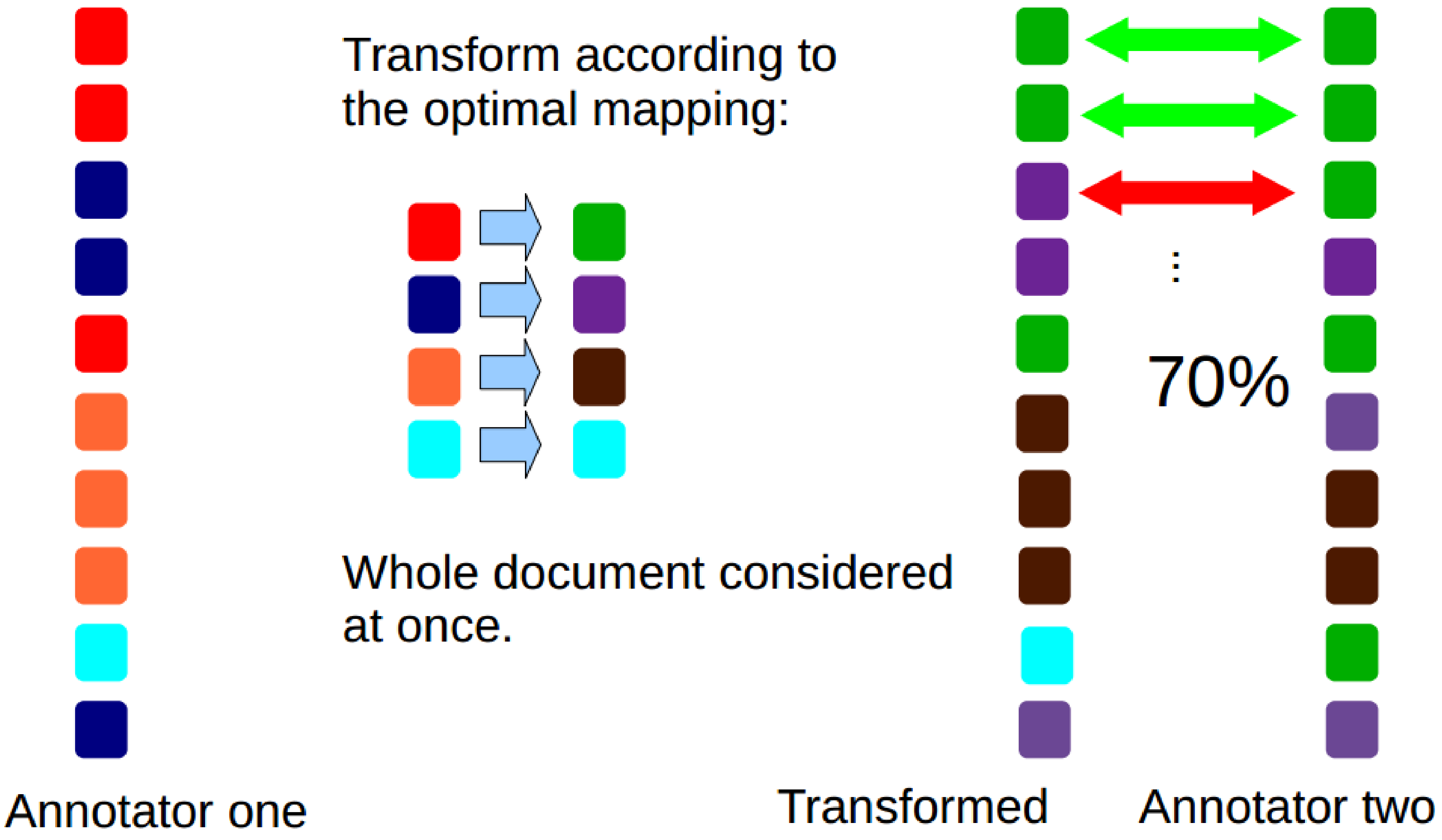
Shen F Metric

$$\mathrm{R} = \frac{n_{ij}}{n_j} \quad \mathrm{P} = \frac{n_{ij}}{n_i} \quad \mathrm{F(i,j)} = \frac{2PR}{P+R}$$

$$\mathrm{F} = \sum_i \frac{n_i}{n} max_j F(i,j)\Big)$$

3. Evaluation Metrics

One-to-One Metric

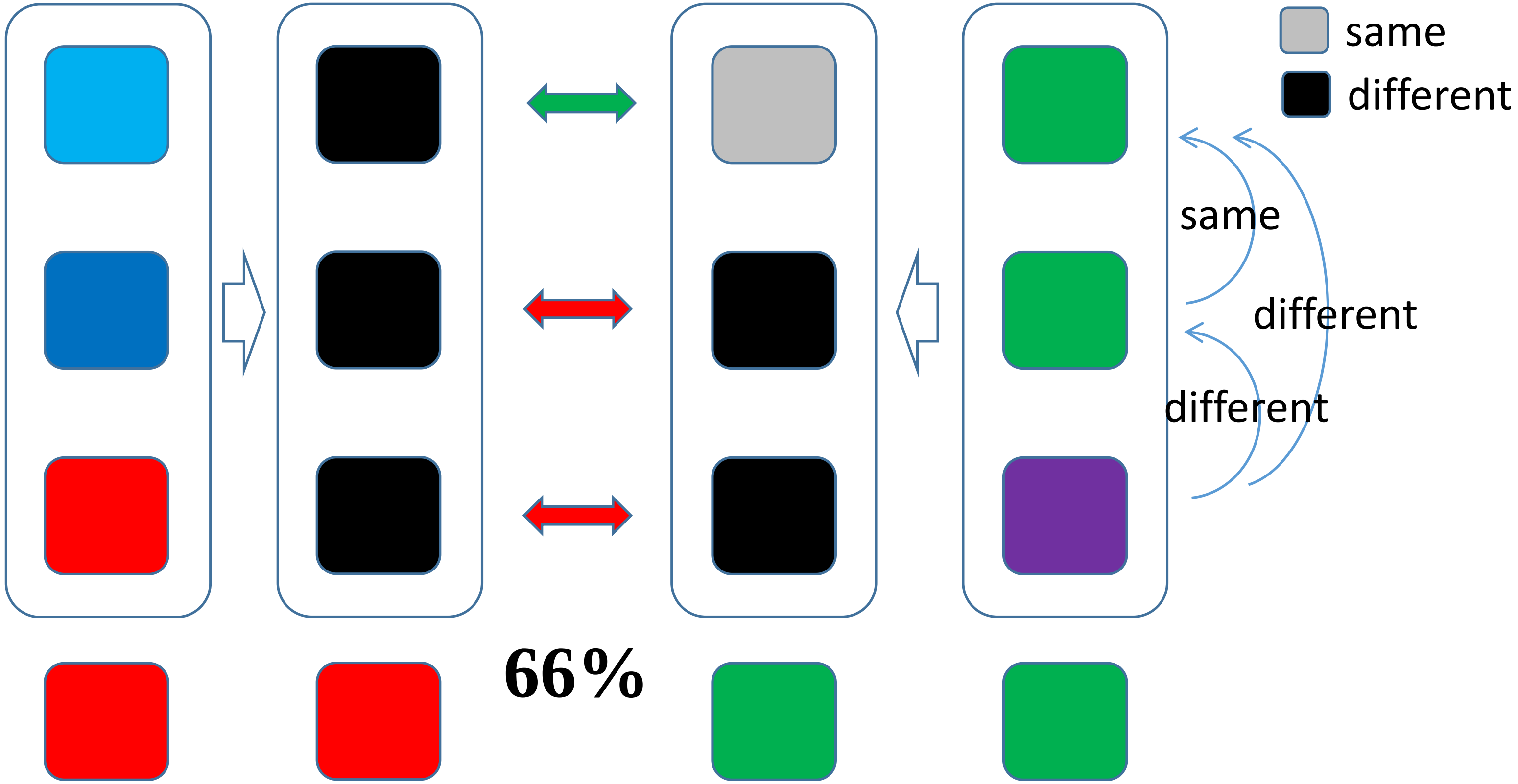


Transform according to the optimal mapping:

Whole document considered at once.

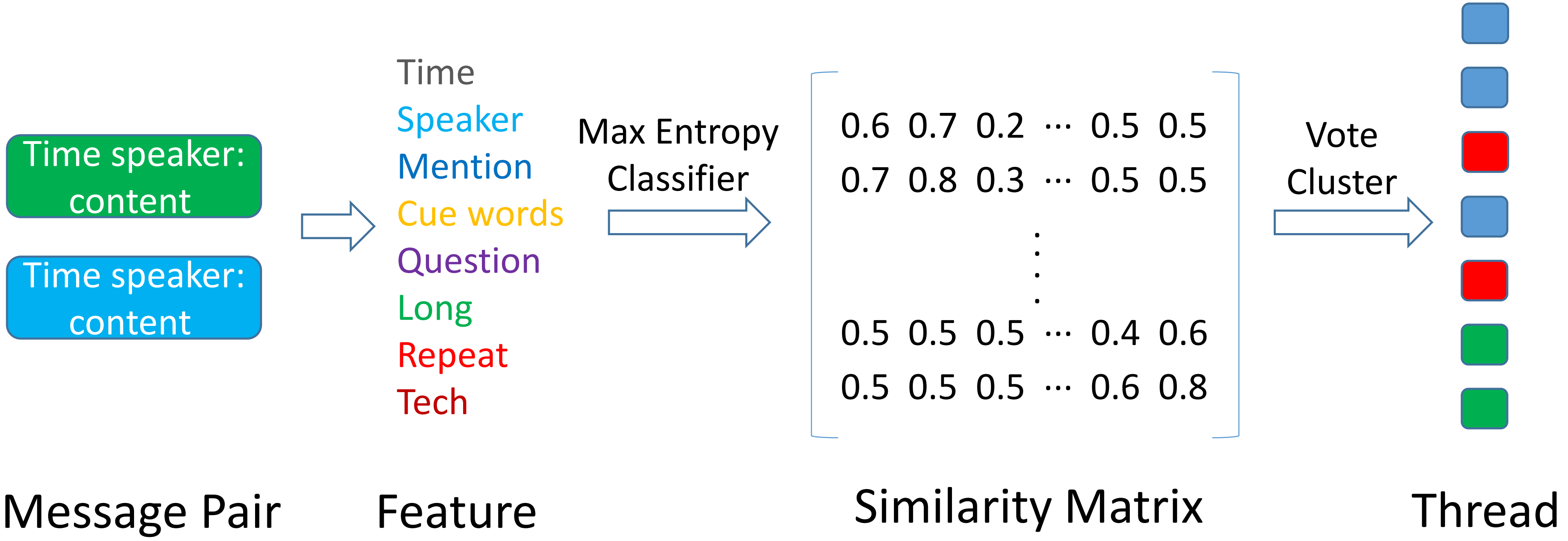
3. Evaluation Metrics

Local Agreement Metric



4. Base model

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In Proceedings of the 46th Annual Meeting of the ACL: HLT (ACL 2008), pages 834–842, Columbus, USA.



Time speaker: content

Time speaker: content

Time

Speaker

Mention

Cue words

Question

Long

Repeat

Tech

Max Entropy Classifier

$$\begin{bmatrix} 0.6 & 0.7 & 0.2 & \cdots & 0.5 & 0.5 \\ 0.7 & 0.8 & 0.3 & \cdots & 0.5 & 0.5 \\ & & & \vdots & & \\ 0.5 & 0.5 & 0.5 & \cdots & 0.4 & 0.6 \\ 0.5 & 0.5 & 0.5 & \cdots & 0.6 & 0.8 \end{bmatrix}$$

Vote Cluster

Message Pair

Feature

Similarity Matrix

Thread

5. Word2vec

5.1 Related Work

Y. Bengio, R. Ducharme, P. Vincent. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155, 2003.

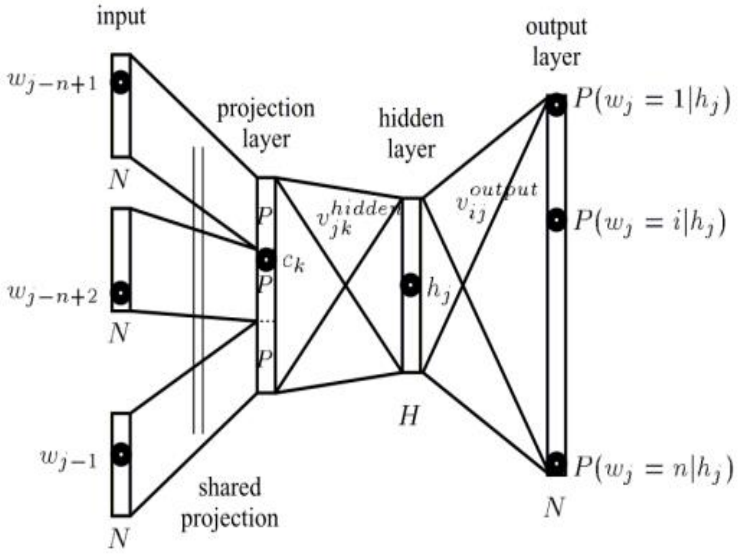
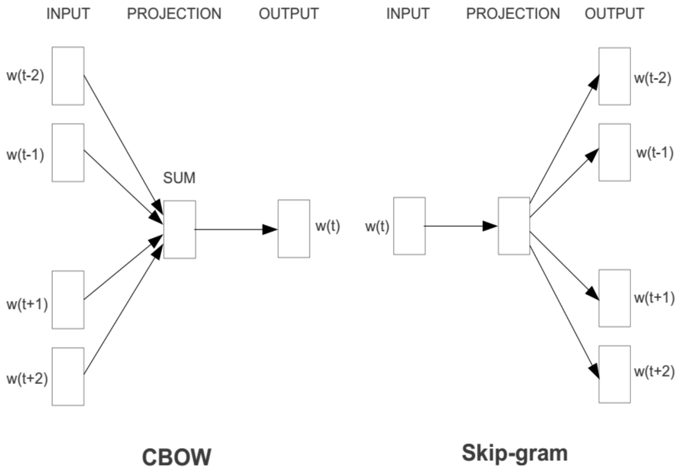


Figure: *Feedforward neural network based LM used by Y. Bengio and H. Schwenk*

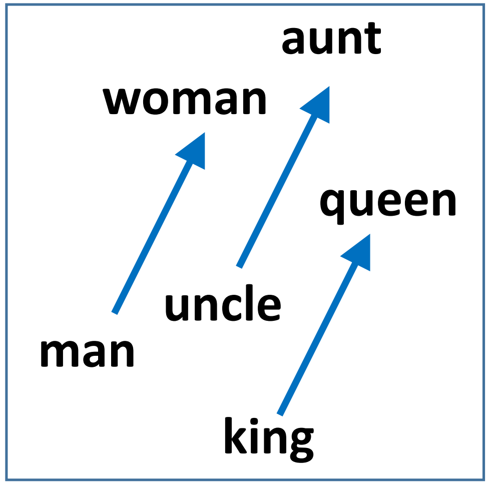


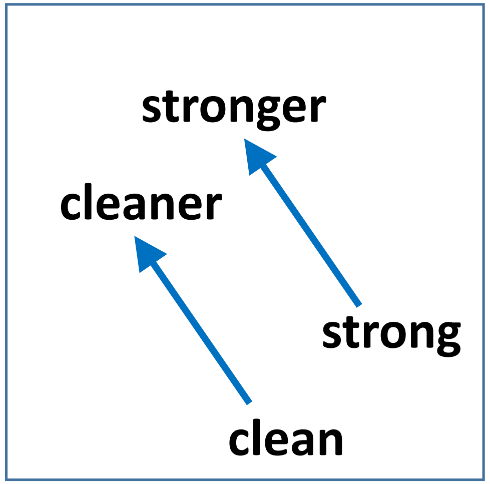
Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013.

5. Word2vec

5.2 Add Word Vector to Features

- Approach1:
  - Pool word vectors to obtain sentence vector
  - Add the similarity between the sentence vector to features
- Approach2:
  - Cluster the word vector in dictionary using K-means
  - Pool the code of words returned by K-means
  - Add the term-by-term product to features





5. Word2vec

5.3 Experiments

Approach 1

	Max F	Mean F	Min F	Max 1-to-1	Mean 1-to-1	Min 1-to-1	Max loc3	Mean loc3	Min loc3
Base Model	56.98	43.91	34.94	54.13	40.63	33.63	75.16	72.75	**70.47**
Average Pooling	57.64	44.68	35.71	51.00	41.79	34.38	74.07	71.45	68.63
Max Pooling	58.57	**45.15**	35.22	51.88	**42.21**	33.75	74.32	71.66	69.05

Max pooling is better than average pooling for Shen F and 1-to-1 metrics.

Approach 2

	Max F	Mean F	Min F	Max 1-to-1	Mean 1-to-1	Min 1-to-1	Max loc3	Mean loc3	Min loc3
Base Model	56.98	43.91	34.94	54.13	40.63	33.63	75.16	72.75	**70.47**
Numclass = 50	57.52	45.01	36.86	50.38	40.73	33.89	73.07	70.32	67.34
Numclass = 75	58.31	45.10	37.48	51.63	41.42	34.88	72.44	70.07	66.92
Numclass = 100	58.11	**45.86**	38.29	51.00	**41.52**	35.50	73.69	70.61	67.42
Numclass = 125	57.07	43.97	35.13	50.88	40.04	32.25	74.11	71.21	68.67
Numclass = 150	57.13	44.60	37.42	51.00	41.06	34.88	72.48	70.26	68.13

The performances of approach 1 and approach 2 are similar.

	Max F	Mean F	Min F	Max 1-to-1	Mean 1-to-1	Min 1-to-1	Max loc3	Mean loc3	Min loc3
Base Model	56.98	43.91	34.94	54.13	40.63	33.63	75.16	72.75	70.47
25d Max Pooling	58.54	45.48	36.08	51.63	42.33	34.13	73.99	71.92	69.55
50d Max Pooling	58.57	45.15	35.22	51.88	42.21	33.75	74.32	71.66	69.05
100d Max Pooling	57.80	44.44	34.43	51.75	41.79	32.75	74.36	72.45	70.34

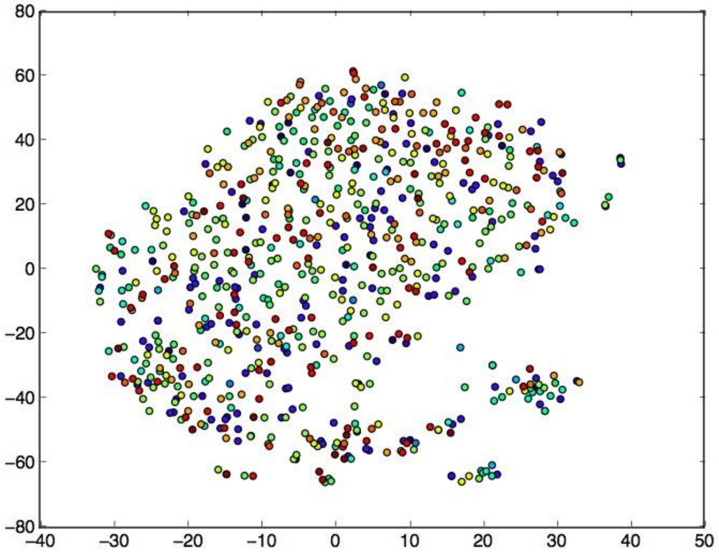
Experiment on Word vector dimension

The dimension of word vectors make little difference for the performance.

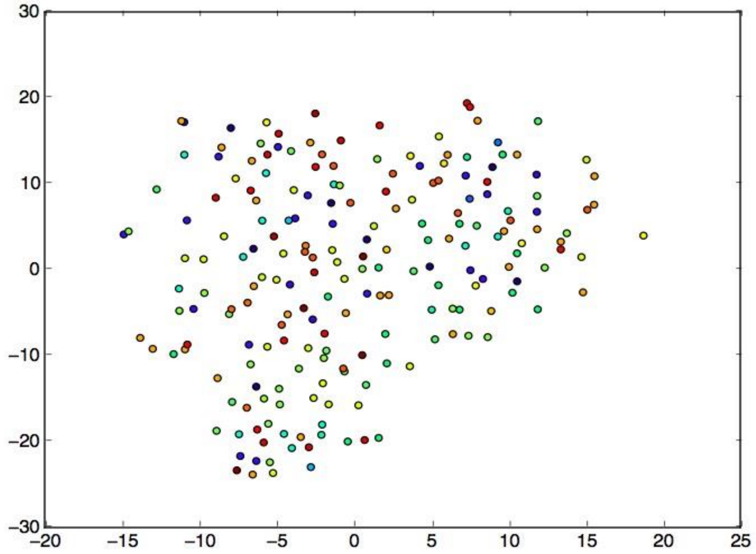
5. Word2vec

5.3 Experiments

Visualization with t-SNE toolkit



All messages

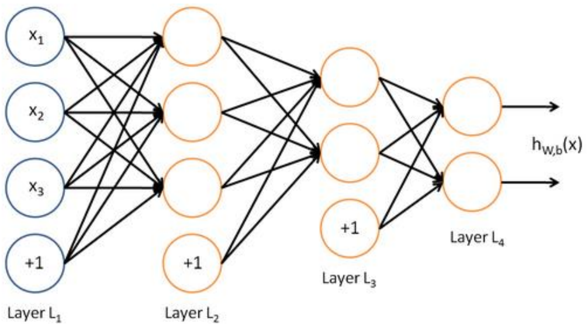


Long messages

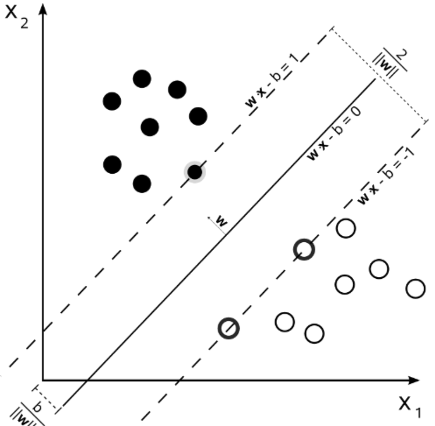
Our Conclusion : The semantic information represented by word vector is little helpful for high level topic detection task.

6. Classifier

6.1 Experiments



Deep Neural Network



SVM

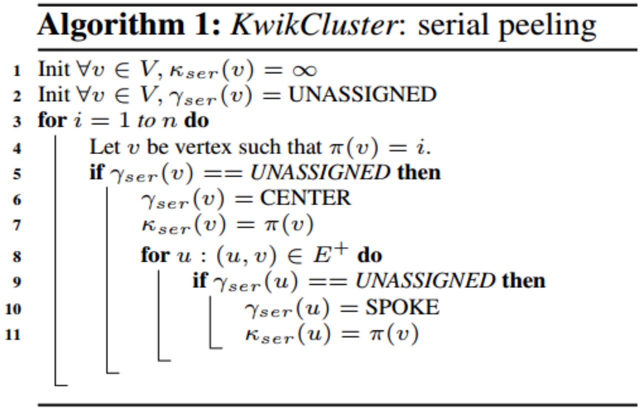
	Max F	Mean F	Min F	Max 1-to-1	Mean 1-to-1	Min 1-to-1	Max loc3	Mean loc3	Min loc3
Base Model	56.98	43.91	34.94	54.13	40.63	33.63	75.16	72.75	**70.47**
SVM	57.08	**47.19**	41.40	49.13	**42.69**	37.63	74.11	69.90	65.12
DNN	60.85	45.36	36.10	55.75	42.40	33.75	74.19	72.72	70.26

Our Conclusion : SVM classifier outperforms the max entropy classifier significantly for Shen F and 1-to-1 metrics, but not as good as max entropy classifier for loc3 metric.

7. Cluster

7.1 Related Work

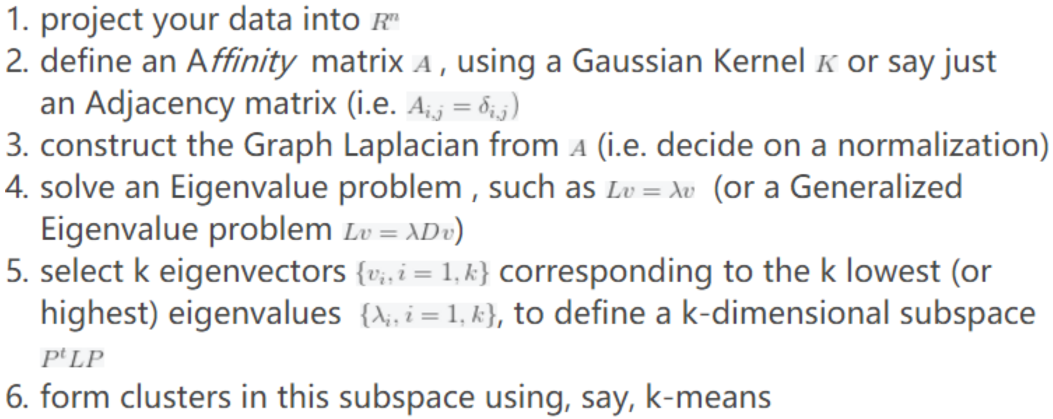
KwikCluster



Algorithm 1: *KwikCluster*: serial peeling

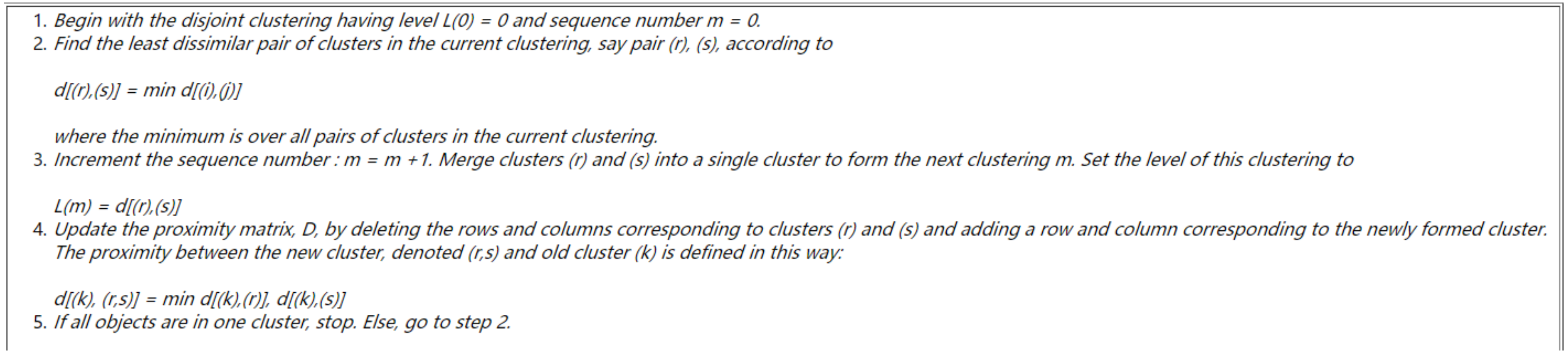
```
Init $\forall v \in V, \kappa_{ser}(v) = \infty$
Init $\forall v \in V, \gamma_{ser}(v) = $ UNASSIGNED
for $i = 1$ to $n$ do
    Let $v$ be vertex such that $\pi(v) = i$.
    if $\gamma_{ser}(v) ==$ UNASSIGNED then
        $\gamma_{ser}(v) = $ CENTER
        $\kappa_{ser}(v) = \pi(v)$
        for $u : (u, v) \in E^+$ do
            if $\gamma_{ser}(u) ==$ UNASSIGNED then
                $\gamma_{ser}(u) = $ SPOKE
                $\kappa_{ser}(u) = \pi(v)$
```

Spectral Clustering



1. project your data into R^n
2. define an A*ffinity* matrix A , using a Gaussian Kernel K or say just an Adjacency matrix (i.e. $A_{i,j} = \delta_{i,j}$)
3. construct the Graph Laplacian from A (i.e. decide on a normalization)
4. solve an Eigenvalue problem , such as $Lv = \lambda v$ (or a Generalized Eigenvalue problem $Lv = \lambda Dv$)
5. select k eigenvectors $\{v_i, i = 1, k\}$ corresponding to the k lowest (or highest) eigenvalues $\{\lambda_i, i = 1, k\}$, to define a k-dimensional subspace $P^t LP$
6. form clusters in this subspace using, say, k-means

Hierarchical Clustering



1. *Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.*
2. *Find the least dissimilar pair of clusters in the current clustering, say pair (r), (s), according to*

   d[(r),(s)] = min d[(i),(j)]

   where the minimum is over all pairs of clusters in the current clustering.
3. *Increment the sequence number : m = m +1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to*

   L(m) = d[(r),(s)]
4. *Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way:*

   d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)]
5. *If all objects are in one cluster, stop. Else, go to step 2.*

7. Cluster

7.2 Experiments

	Max F	Mean F	Min F	Max 1-to-1	Mean 1-to-1	Min 1-to-1	Max loc3	Mean loc3	Min loc3
Base Model	56.98	43.91	34.94	54.13	40.63	33.63	75.16	72.75	70.47
KwikCluster	51.64	43.18	28.45	46.63	37.54	23.63	71.39	68.41	64.70
Spectral Clustering	51.64	39.88	22.94	45.13	34.35	21.13	70.18	67.66	64.03
Hierarchical Clustering	44.06	19.30	9.29	52.13	23.19	10.88	51.61	43.33	35.09

- KwikCluster is competitive to vote greedy algorithm. It is worth noting that a parallel variant of KwikCluster is proposed in [5]. So it is scalable for big data.
- Spectral Clustering is not suitable for this task.
- Although hierarchical clustering doesn’t work well, it inspire the research in evaluation method

8. Evaluation Method

8.1 Motivation

- Everyone has his/her own granularity.
  - Fig1 shows that the granularity between annotators is very different
- Original evaluation method fix parameter tv which controls granularity to 0.5. Granularity difference will influence the evaluation of algorithm
  - Fig2 shows when tv changes the performance changes a lot

Fig1

Annotation	test-0.annot	test-1.annot	test-2.annot	test-3.annot	test-4.annot	test-5.annot
Thread number	71	129	93	71	51	79

Fig2

	0.3	0.4	0.5	0.6	0.7	0.8	0.9
test-0.annot	19.30	36.64	44.71	48.77	54.97	57.63	51.48
test-1.annot	14.54	29.19	34.94	40.68	47.03	49.28	55.73
test-2.annot	14.54	29.19	34.94	40.68	47.03	49.28	55.73
test-3.annot	17.43	33.33	40.9	45.67	52.17	53.65	49.59
test-4.annot	22.32	36.38	39.42	42.48	47.29	48.12	44.97
test-5.annot	53.81	61.51	56.98	55.86	42.39	43.99	33.19

8. Evaluation Method

8.2 New Evaluation Method

- Vary the parameter tv which controls granularity within certain range.
- Find the parameter tv with which algorithm has the best performance.
- Average the best metrics computed against each annotation

$$\text{Shen F} = \frac{1}{n}\sum_{i=1}^{n} max_{tv} F_i$$

$$\text{1-to-1} = \frac{1}{n}\sum_{i=1}^{n} max_{tv} \text{1−to−1}_{\ i}$$

$$loc_3 = \frac{1}{n}\sum_{i=1}^{n} max_{tv} {loc_3}_i$$

8. Evaluation Method

8.3 Experiments

	Shen F var	1-to-1 var	Loc3 var
tv = 0.5	48.03	**28.43**	3.76
Best tv	**18.13**	40.55	**1.83**

All features
Max entropy classifier

	Shen F var	1-to-1 var	Loc3 var
tv = 0.5	**12.04**	**4.90**	21.02
Best tv	18.07	59.70	**5.32**

Without mention feature
Max entropy classifier

	Shen F var	1-to-1 var	Loc3 var
tv = 0.5	40.83	**39.09**	3.42
Best tv	**22.57**	49.11	**2.44**

Without speaker feature
Max entropy classifier

	Shen F var	1-to-1 var	Loc3 var
tv = 0.5	54.37	**33.65**	3.37
Best tv	**20.47**	46.28	**2.27**

Without time feature
Max entropy classifier

	Shen F var	1-to-1 var	Loc3 var
tv = 0.5	57.39	**42.93**	2.75
Best tv	**38.17**	47.08	**2.07**

All features
SVM classifier

Our Conclusion : New evaluation method can reduce the variation of the metrics so that it is more consistent for algorithm evaluation

Reference

- [1] Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In Proceedings of the 46th Annual Meeting of the ACL: HLT (ACL 2008), pages 834–842, Columbus, USA.
- [2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.
- [3] Y. Bengio, R. Ducharme, P. Vincent. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155, 2003.
- [4] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[J]. Advances in neural information processing systems, 2002, 2: 849-856.
- [5] Pan X, Papailiopoulos D, Oymak S, et al. Parallel Correlation Clustering on Big Graphs[C]//Advances in Neural Information Processing Systems. 2015: 82-90.

Thank you for your attention！

Stay Hungry, Stay Foolish.

wangyang@cslt.riit.tsinghua.edu.cn