Temporal Proximity (TP). Re-mention of adjacent temporal information may strengthen event continuousness and raise significance, but huge time gap indicates separate events. Given $\Delta t = |t_n - t|$ where t_n is the new time contained in sentence s_t and t is from s_h, T_D as the time span of D, temporal proximity $f_t(x) = e^{-\alpha \times \frac{\Delta t}{T_D}} \times f_d(x - ||s_t||)$.

Named Entities (NE). Sentence with named entities (s_e) might indicate strong relevance if entities are connected by existing knowledge databases (e.g. WordNet or Wikipedia), but [2] assumed equal distance for all adjacent entities in hierarchical taxonomy structures. Leaf/lower level entities should be closer than general concepts from higher levels. Consider a fragment, <health [food safety, public health organization(Centers for Disease Control, World Health Organization)]>, (CDC, WHO) are closer than (food safety, public health organization). We model synonyms, hyponyms and hypernyms into entity distance. We assign a distance weight (w_e) to every entity $w_e = 1 + \sum_{e_k \in H(e)} w_{e_k}$ where $H(e)$ is the hyponym set of entity e. The distance from a hyernym e to one of its hyponym e_k is defined as:

$$dist = \sum_{e_k \in H(e)} \frac{w_{e_k}}{|H(e)|}. \qquad (3)$$

The weight of leaf node is set as 1. *dist* and *weight* are measured separately and penalization costs more for category entities. Entity influence $f_e(x) = e^{-\beta \times dist} \times f_d(x - ||s_e||)$.

α, β are scaling factors. $f_d(x)$, $f_e(x)$, $f_t(x)$ affect sentence significance separately and there are more than one s_e or s_t in S. For snippet completeness we choose the maximum $f_e(x)$ and $f_t(x)$ and take the arithmetic average of the three.

Conjunctive Indicators (CI). Conjunctions such as "however", "so", etc. reflect the author's intention of a semantic bridge between the adjacent sentences, which raises sentence significance. For the sentence with these conjunctive indicators, we assume it shares the same significance with its neighboring sentence prior to it. The conjunctive influence is local and not accumulative to following texts.

$$sig(x) = sig(x-1) \quad \text{if } (s_x \cap s_{x-1}) \subseteq \text{CI}. \qquad (4)$$

Layout Presentation (LP). The visual structure of the news article in the webpage can give some clues to the event atoms, since writing style implies event principles as well.

- *Line break.* When meet the tag of <br> or <p>, the line break as the author's intention of topic drifting.
- *Visual Elements.* An inserted image, table or hyperlink (<img>, <a>, etc.) indicates similar effect as line breaks due to news writing style.

The effects of line break and visual elements are accumulative. After τ visual changes, the probability drops by $\prod_\tau (1 - r_i)$. r_i are not equal due to specific contexts but for simplicity we assume they are all r. Hence final $sig(.)$ is:

$$sig(x) = (f_d(x) + max\{f_e(x)\} + max\{f_t(x)\}) \times (1 - r)^\tau / 3 \qquad (5)$$

Combining Significance. Each sentence in snippet affects following sentences, either increasing or decreasing the significance. We apply $sig(.)$ in Equation (1) and obtain a weighted relevance score from all sentence pairs between s_p and sentences in the expanding snippet S. We add s_p into S when relevance exceeds a threshold.

$$p(s_p|LM(S)) = \left(\prod_{w \in s_p} \frac{\sum_{s_i \in S} sig(s_i) \cdot tf(w, s_i) + \lambda}{(1+\lambda) \cdot \sum_{s_i \in S} sig(s_i) \cdot |s_i|} \right)^{\frac{1}{|s_p|}} \qquad (6)$$

3. EXPERIMENTS

In a 10-fold cross validation manner, we test our proposed approaches on a corpus of 1000 webpages from the *Xinhua News* website. There are on average 1.893 snippets per news document and for all snippets, $\mu = 6.97$, $\sigma = 2.11$. Golden standards are created by human annotators. α, β, r are set experimentally at 0.6, 0.5, 0.174 correspondingly. We stick to the *precision/recall* evaluation metrics in [2]. Figure 1 shows the experiment results of semantic relevance (SeRel) and weighted semantic relevance (WSeRel) compared with TextTiling proposed in [1], TTM and LGM proposed in [2]. The perfromance of different features is shown in Figure 2.

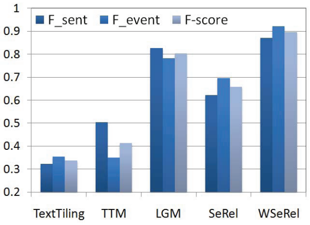


Figure 1: Performance

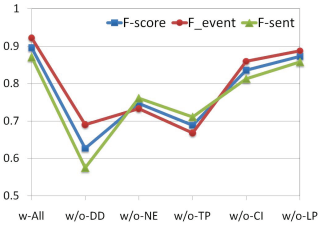
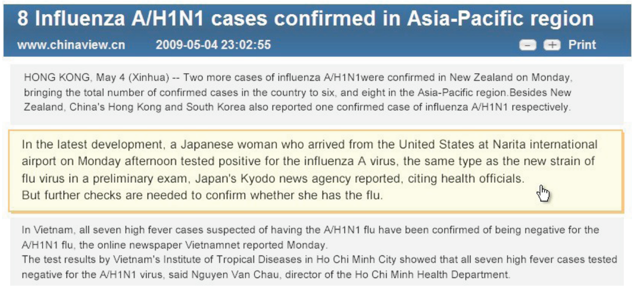


Figure 2: Features

WSeRel generally outperforms others. TextTiling shows significant weakness because it is not event-oriented. The contribution of **significance** is obvious (+26.56%) by comparing WSeRel with SeRel. DD is the most essential for snippet expansion. TP, NE, CI are also necessary. LP seems not to perform well due to misleading line breaks and visual noises. We present a system demonstration snapshot.



8 Influenza A/H1N1 cases confirmed in Asia-Pacific region

www.chinaview.cn 2009-05-04 23:02:55

HONG KONG, May 4 (Xinhua) -- Two more cases of influenza A/H1N1were confirmed in New Zealand on Monday, bringing the total number of confirmed cases in the country to six, and eight in the Asia-Pacific region.Besides New Zealand, China's Hong Kong and South Korea also reported one confirmed case of influenza A/H1N1 respectively.

In the latest development, a Japanese woman who arrived from the United States at Narita international airport on Monday afternoon tested positive for the influenza A virus, the same type as the new strain of flu virus in a preliminary exam, Japan's Kyodo news agency reported, citing health officials.
But further checks are needed to confirm whether she has the flu.

In Vietnam, all seven high fever cases suspected of having the A/H1N1 flu have been confirmed of being negative for the A/H1N1 flu, the online newspaper Vietnamnet reported Monday.
The test results by Vietnam's Institute of Tropical Diseases in Ho Chi Minh City showed that all seven high fever cases tested negative for the A/H1N1 virus, said Nguyen Van Chau, director of the Ho Chi Minh Health Department.

Figure 3: Fine-grained news digestion system demo.

4. CONCLUSIONS

We describe a fine-grained news digestion framework of ESE, utilizing semantic, syntactic and visual features. ESE is an on-going infrastructure work facilitating other researches. We show that our approach outperforms rival methods.

5. ACKNOWLEDGMENTS

This work is partially supported by NSFC with Grant No.60933004, 61050009 and 61073081.

6. REFERENCES

[1] M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, 1997.

[2] R. Yan, Y. Li, Y. Zhang, and X. Li. Event recognition from news webpages through latent ingredients extraction. In *AIRS '10*, pages 490–501.