Tracking Information Flow between Primary and Secondary News Sources

Will Radford [†‡] **Ben Hachey** [‡◇] **James R. Curran** [†‡] **Maria Milosavljevic** [◇]

School of Information Technologies[†]
University of Sydney
NSW 2006, Australia

Capital Markets CRC[‡]
55 Harrington Street
NSW 2000, Australia

Centre for Language Technology[◇]
Macquarie University
NSW 2109, Australia

{wradford,james}@it.usyd.edu.au bhachey@cmcrc.com mariam@ics.mq.edu.au

Abstract

Tracking information flow (IFLOW) is crucial to understanding the evolution of news stories. We present analysis and experiments for IFLOW between company announcements and newswire. Error analysis shows that many FPs are annotation errors and many FNs are due to coarse-grained document-level modelling. Experiments show that document meta-data features (e.g., category, length, timing) improve f-scores relative to upper bound by 23%.

1 Introduction

Tracking IFLOW between primary and secondary news sources provides insight into the contribution of participants and the role of sources. In finance, being alert and responsive to the nature of incoming information (e.g., novelty, price sensitivity) is central to successful trading (Zaheer and Zaheer, 1997). Traders need tools that flag price-sensitive information in a high-volume news feed. IFLOW is central to market surveillance, where unusual market activity (e.g., abnormal changes in trading price or volume) is linked to explanations in the information ecosystem (Milosavljevic et al., 2009).

In Australia, the Australian Securities Exchange (ASX) is the official syndicator of information that might affect a company's share price. Subsequently, a variety of secondary sources (e.g., news media, blogs, forums) repackage this information. We focus on the relationship between ASX company announcements and Reuters newswire, which filters and aggregates the key details from company announcements in near-real time.

2 Preliminary Results

We define IFLOW for capital markets as *a pair of documents where one repeats price-sensitive information from the other* (Radford et al., 2009). Pairs of ASX announcements and Reuters NewsScope Archive (RNA) stories covering the same company and released within a week of one another are manually annotated for presence or absence of IFLOW. These are used to train MEGAM (Daumé III, 2004) maximum entropy models for identifying IFLOW. Textual features include set-theoretic bags of word unigrams and bigrams over the document text and titles. Text, title and numeric token similarity scores (Metzler et al., 2005) provide a more general notion of similarity. The precision of numeric tokens is also represented. Counts of matched sentences and longest common sub-sequences capture longer units of reused text. Temporal features model the news cycle and news source responsiveness.

In development experiments (ten-fold cross validation, 30,249 ASX-RNA pairs), the system identifies IFLOW pairs at 89.5% f-score (Radford et al., 2009). In evaluation experiments (held-out test set, 1,621 ASX-RNA pairs), it achieves 76.6% f-score, significantly better than a text-only baseline (62.5%) and 10% less than the human upper bound (86.4%).

3 Error analysis

We engaged finance students (fourth-year or higher) to examine the 20 false positive (FP) errors with the highest IFLOW probabilities and the 20 false negative (FN) errors with the lowest IFLOW probabilities. Table 1 shows the resulting reassessment of the

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 29–30,
Los Angeles, California, June 2010.

Error	Correct	Incorrect	Ambiguous
FP	4 (20%)	15 (75%)	1 (5%)
FN	15 (75%)	4 (20%)	1 (5%)

Table 1: Analysis of original annotation correctness.

original IFLOW annotation. For FPs, 75% were determined to have been incorrectly annotated as absent of IFLOW. This is not unexpected since IFLOW can be based on small details (e.g., '$2.45m profit') which are easily missed by annotators. This suggests that the system's actual precision may be higher than 90.9%. Mis-annotation is less common for FNs (20%). However, the proportion of DIGEST documents (those that report on multiple events) is much higher for FNs (75% compared to 30% for FPs). It is likely that legitimate textual similarity is lost in the noise of the irrelevant content.

4 Document Metadata Features

We add new features that take advantage of categorisation information in the source metadata. These include ASX tags for price sensitivity, ASX and RNA type tags and journalist revision comments embedded in RNA stories. These features model differences in IFLOW between document types (e.g., periodic reports are more likely to be reported than a dividend rate announcement). A feature representing the length of each ASX-RNA document is also included. We also add detail to the temporal features, including the day and month the announcement was released, as well as whether the announcement and story were released on the same day.

The metadata features lead to significantly better f-score in development experiments (Table 2). Subtractive feature analysis suggests that the document type and length features are effective ($p<0.05$) but the detailed temporal features are not. The revision comments are borderline ($p=0.051$). In Table 3, the metadata features improve the f-score by 23% over Radford et al. (2009) with respect to the upper bound, but the difference is not significant. The different precision-recall balance between experiments is consistent with Section 3.

5 Discussion and Future Work

We have developed a dataset for IFLOW in the context of financial text mining and demonstrated it is a feasible task using simple approaches. Future work will involve more advanced models. First, we will consider *sub-document* analysis, as suggested by the DIGEST FNs in the error analysis. This will also enable tools that highlight specific types of contribution (e.g., adding background context, novel analysis) within secondary sources. Furthermore, the wider IFLOW ecosystem includes other sources (e.g., bloggers, forum contributors) that should be analysed for leading and lagging indicators. Finally, a number of specific applications might serve as extrinsic evaluations of the IFLOW task. These include de-duplicating and aggregating information feeds and automatically attributing reported content to a source story.

Features	P (%)	R (%)	F (%)
Radford et al. (2009)	90.9	88.1	89.5
+ Metadata Features	**91.1**	$\star\star$**89.3**	$\star$**90.2**

Table 2: Precision (P), recall (R) and f-score (F) for development experiments ($\star$: $p<0.05$, $\star\star$: $p<0.01$).

Features	P (%)	R (%)	F (%)
Text-only Baseline	80.0	51.3	62.5
Radford et al. (2009)	84.5	70.1	76.6
+ Metadata Features	**86.3**	**72.6**	**78.9**
Human Upper Bound	88.9	85.1	86.4

Table 3: P, R and F for evaluation experiments.

References

Hal Daumé III. 2004. Notes on CG and LM-BFGS optimization of logistic regression. `http://hal3.name/docs/daume04cg-bfgs.pdf`.

Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel. 2005. Similarity measures for tracking information flow. In *Proc. CIKM*, pages 517–524.

Maria Milosavljevic, Jean-Yves Delort, Ben Hachey, Bavani Arunasalam, Will Radford, and James R. Curran. 2009. Automating financial surveillance. In *Proceedings of the Workshop on Mining User-Generated Content for Security*.

Will Radford, Ben Hachey, James R. Curran, and Maria Milosavljevic. 2009. Tracking information flow in financial text. In *Proc. ALTA*, pages 11–19.

Akbar Zaheer and Srilata Zaheer. 1997. Catching the wave: alertness, responsiveness, and market influence in global electronic networks. *Management Science*, 43(11):1493–1509.