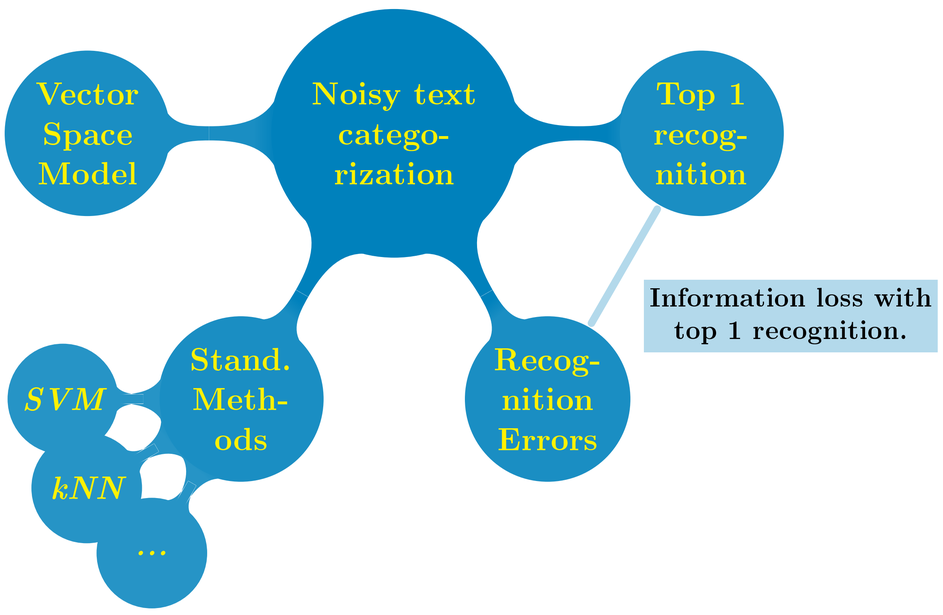
Using top n Recognition Candidates to Categorize On-line Handwritten Documents

Sebastián Peña Saldarriaga[†], Emmanuel Morin[†] et Christian Viard-Gaudin[‡]
{sebastian.pena-saldarriaga, emmanuel.morin, christian.viard-gaudin}@univ-nantes.fr
(†) LINA — Université de Nantes, (‡) IRCCyN — École Polytechnique de l'Université de Nantes

1 Introduction

- **Archival & retrieval of on-line handwriting**
  - Particular interest for text categorization (TC)
  - TC attempts to derive information from text
  - Recognition is a necessary effort



2 Intuitive Idea

- **Use top n (n > 1) recognition candidates**
  - Greater probability of having the correct word

n	1	5	10	15
Word level rec.	22.08%	15.73%	13.41%	12.04%
Char. level rec.	52.48%	37.28%	33.09%	31.02%

Recognition rates for different n

However, the text is flooded with false occurrences of words. We then redefine the standard $tf \times idf$ weighting scheme.

3 Weighting top n candidates

- The **weight** w of a term i is given by

$$w_i = \frac{ctf(i) \times idf(i)}{\sqrt{\sum_{j=1}^{M}(ctf(j) \times idf(i))^2}}$$

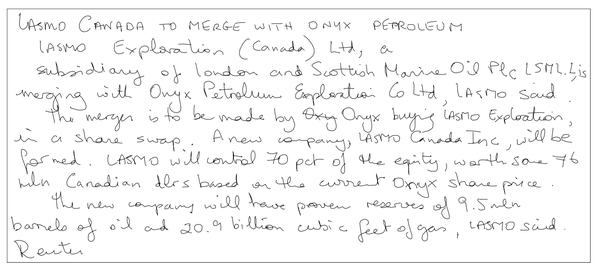
- & the **candidate-term frequency** (ctf) by

$$ctf(i) = \sum_{n=1}^{N} p_n(i)$$

The sum of the probabilities of i in the N candidate lists in which i occurs

4 Data & Experiments

- **On-line handwritten data**

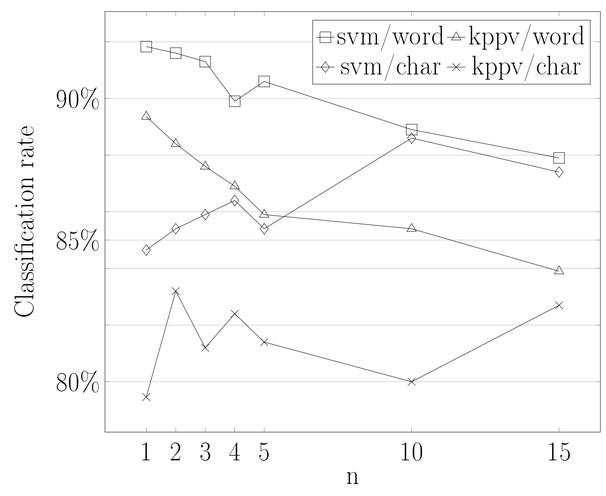

LASMO CANADA TO MERGE WITH ONYX PETROLEUM
LASMO Exploration (Canada) Ltd, a subsidiary of London and Scottish Marine Oil Plc LSML.L, is merging with Onyx Petroleum Exploration Co Ltd, LASMO said.
The merger is to be made by Oxy Onyx buying LASMO Exploration, in a share swap. A new company, LASMO Canada Inc, will be formed. LASMO will control 70 pct of the equity, worth some 76 mln Canadian dlrs based on the current Onyx share price.
The new company will have proven reserves of 9.5 mln barrels of oil and 20.9 billion cubic feet of gas, LASMO said.
Reuter

- Reuters-21578 corpus
- 2,000 samples
- 10 categories

- **Experiments**
  - Comparing *kNN* & *SVM*
  - With *word* & *char.* level recognition
  - With 1 or n recognition candidates

5 Results

- **Word level rec.**
  - Accuracy *decreases*
  - ... as n increases
- **Char. level rec.**
  - Accuracy *improves*
  - ... $\forall\ n > 1$
  - With *both* algorithms



6 Conclusion

- **What has been accomplished ?**
  - Redefinition of the $tf \times idf$ weighting
  - Based on probabilities of recognition candidates

✓ The good

A simple idea that yields interesting results on heavily degraded texts with two different classifiers.

✗ The bad

However, it is uneffective with word level recognition. Needs further experimental validation.

- **What's next ?**
  - Thresholding/rejection strategies on candidates
  - Track & type recognition errors

Acknowledgments

This work is funded by La Région Pays de la Loire under the MILES Project and by The French National Research Agency grant number ANR-06-TLOG-009.