PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 169–179

# FaDA: Fast Document Aligner using Word Embedding

Pintu Lohar, Debasis Ganguly, Haithem Afli, Andy Way, Gareth J.F. Jones

ADAPT Centre
School of computing
Dublin City University
Dublin, Ireland

**Abstract**

FaDA[1] is a free/open-source tool for aligning multilingual documents. It employs a novel crosslingual information retrieval (CLIR)-based document-alignment algorithm involving the distances between embedded word vectors in combination with the word overlap between the source-language and the target-language documents. In this approach, we initially construct a pseudo-query from a source-language document. We then represent the target-language documents and the pseudo-query as word vectors to find the average similarity measure between them. This word vector-based similarity measure is then combined with the term overlap-based similarity. Our initial experiments show that s standard Statistical Machine Translation (SMT)-based approach is outperformed by our CLIR-based approach in finding the correct alignment pairs. In addition to this, subsequent experiments with the word vector-based method show further improvements in the performance of the system.

## 1. Introduction

A crosslingual document alignment system aims at efficiently extracting likely candidates of aligned documents from a comparable corpus in two or more different languages. Such a system needs to be effectively applied to a large collection of documents. As an alternative approach, a state-of-the-art machine translation (MT) system (such as Moses, Koehn et al., (2007)) can be used for this purpose by translateing every source-language document with an aim of representing all the documents in the

[1] Available at `https://github.com/gdebasis/cldocalign/`

 Corresponding author: haithem.afli@adaptcentre.ie
Cite as: Pintu Lohar, Debasis Ganguly, Haithem Afli, Andy Way, Gareth J.F. Jones. FaDA: Fast Document Aligner using Word Embedding. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 169–179.
doi: 10.1515/pralin-2016-0016.

same vocabulary space. This in turn facilitates the computation of the text similarity between the source-language and the target-language documents. However, this approach is rather impractical when applied to a large collection of bilingual documents, because of the computational overhead of translating the whole collection of source-language documents into the target language.

To overcome this problem, we propose to apply an inverted index-based cross-language information retrieval (CLIR) method which does not require the translation of documents. As such, the CLIR apporach results in much reduction computation compared to the MT-based method. Hence we refer to our tool using the CLIR approach as the *Fast document aligner (FaDA)*. Our FaDA system works as follows. Firstly, a pseudo-query is constructed from a source-language document and is then translated with the help of a dictionary (obtained with the help of a standard word-alignment algorithm (Brown et al., 1993) using a parallel corpus). The pseudo-query is comprised of the representative terms of the source-language document. Secondly, the resulting translated query is used to extract a ranked list of documents from the target-language collection. The document with the highest similarity score is considered as the most likely candidate alignment with the source-language document.

In addition to adopted a standard CLIR query-document comparison, the FaDA systems explores the use of a word-vector embedding approach with the aim of building a semantic matching model in seeks to improve the performance of the alignment system. The word-vector embedding comparison method is based on the relative distance between the embedded word vectors that can be estimated by a method such as 'word2vec' (Mikolov et al., 2013). This is learned by a recurrent neural network (RNN)-based approach on a large volume of text. It is observed that the inner product between the vector representation of two words $u$ and $v$ is high if $v$ is likely to occur in the context of $u$, and low otherwise. For example, the vectors of the words 'child' and 'childhood' appear in similar contexts and so are considered to be close to each other. FaDA combines a standard text-based measure of the vocabulary overlap between document pairs, with the distances between the constituent word vectors of the candidate document pairs in our CLIR-based system.

The remainder of the paper is organized as follows. In Section 2, we provide a literature survey of the problem of crosslingual document alignment. In Section 3, the overall system architecture of *FaDA* is described. In Section 4, we describe our experimental investigation. The evaluation of the system is explained in Section 5. Finally, we conclude and suggest possible future work in Section 6.

## 2. Related Work

There is a plethora of existing research on discovering similar sentences from comparable corpora in order to augment parallel data collections. Additionally, there is also existing work using the Web as a comparable corpus in document alignment. For example, Zhao and Vogel (2002) mine parallel sentences from a bilingual compa-

rable news collection collected from the Web, while Resnik and Smith (2003) propose a web-mining-based system, called STRAND, and show that their approach is able to find large numbers of similar document pairs. Bitextor[2] and ILSPFC[3] follow similar web-based methods to extract monolingual/multilingual comparable documents from multilingual websites.

Yang and Li (2003) present an alignment method at different levels (title, word and character) based on dynamic programming (DP) to identify document pairs in an English-Chinese corpus collected from the Web, by applying the longest common subsequence to find the most reliable Chinese translation of an English word. Utiyama and Isahara (2003) use CLIR and DP to extract sentences from an English-Japanese comparable corpus. They identify similar article pairs, consider them as parallel texts, and then align the sentences using a sentence-pair similarity score and use DP to find the least-cost alignment over the document pair.

Munteanu and Marcu (2005) use a bilingual lexicon to translate the words of a source-language sentence to query a database in order to find the matching translations. The work proposed in Afli et al. (2016) shows that it is possible to extract only 20% of the true parallel data from a collection of sentences with 1.9M tokens by employing an automated approach.

The most similar work to our approach is described in Roy et al. (2016). In this documents and queries are represented as sets of word vectors, similarity measure between these sets calculated, and then combine with IR-based similarities for document ranking.

## 3. System architecture of FaDA

The overall architecture of FaDA comprises two components; (i) the CLIR-based system, and (ii) the word-vector embedding system.

### 3.1. CLIR-based system

The system diagram of our CLIR-based system is shown in Figure (1). The source-language and the target-language documents are first indexed, then each of the indexed source-language documents is used to construct a pseudo-query. However, we do not use all the terms from a source-language document to construct the pseudo-query because very long results in a very slow retrieval process. Moreover, it is more likely that a long query will contain many 'outlier' terms which are not related to the core topic of the document, thus reducing the retrieval effectiveness. Therefore, we use only a fraction of the constituent terms to construct the pseudo-query, which are considered to be suitably representative of the document.

[2] http://bitextor.sourceforge.net/

[3] http://nlp.ilsp.gr/redmine/projects/

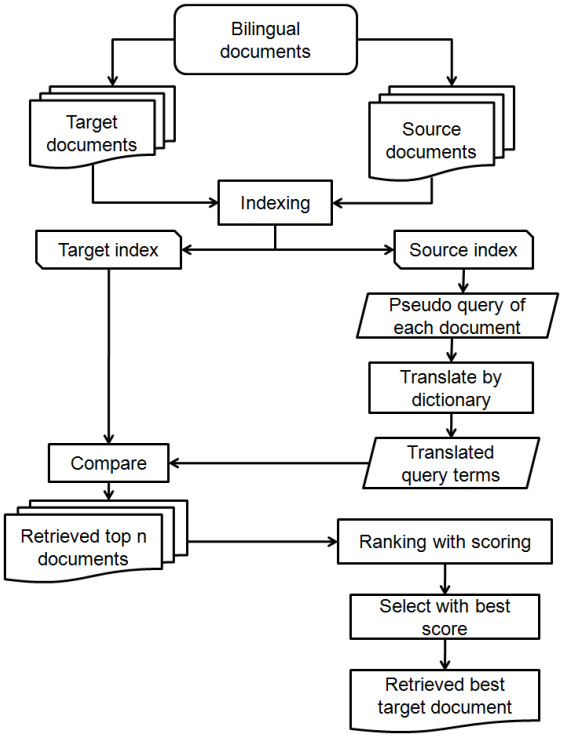


*Figure 1. Architecture of the CLIR-based system*

To select the terms to include in the pseudo-query we use the score shown in Equation (1), where $\mathrm{tf}(t, d)$ denotes the term frequency of a term $t$ in document $d$, $\mathrm{len}(d)$ denotes the length of $d$, and $N$ and $\mathrm{df}(t)$ denote the total number of documents and the number of documents in which $t$ occurs, respectively. Furthermore, $\tau(t, d)$ represents the term-selection score and is a linear combination of the normalized term frequency of a term $t$ in document $d$, and the inverse document frequency (idf) of the term.

$$\tau(t, d) = \lambda \frac{\mathrm{tf}(t, d)}{\mathrm{len}(d)} + (1 - \lambda) \log\left(\frac{N}{\mathrm{df}(t)}\right) \quad (1)$$

It is obvious that in Equation (1) the terms that are frequent in a document $d$ and the terms that are relatively less frequent in the collection are prioritized. The parameter $\lambda$ controls the relative importance of the $\mathrm{tf}$ and the $\mathrm{idf}$ factors. Using this function, each term in $d$ is associated with a score. This list of terms is sorted in de-

creasing order of this score. Finally, a fraction σ (between 0 and 1) is selected from this sorted list to construct the pseudo-query from d. Subsequently, the query terms are translated by a source-to-target dictionary, and the translated query terms are then compared with the indexed target-language documents. After comparison, the top-n documents are extracted and ranked using the scoring method in Equation (3), which is explained in Section 3.2.1. Finally, to select the best candidate for the alignment, we choose the target-language document with the highest score.

### 3.2. Word-vector embedding-based system

In addition to the CLIR framework described in Section 3.1, we also use the vector embedding of words and incorporate them with the CLIR-based approach in order to estimate the semantic similarity between the source-language and the target-language documents. This word-embedding approach facilitates the formation of "bag-of-vectors" (BoV) which helps to express a document as a set of words with one or more clusters of words where each cluster represents a topic of the document.

Let the BoW representation of a document d be $W_d = \{w_i\}_{i=1}^{|d|}$, where $|d|$ is the number of unique words in d and $w_i$ is the $i^{th}$ word. The BoV representation of d is the set $V_d = \{x_i\}_{i=1}^{|d|}$, where $x_i \in \mathbb{R}^p$ is the vector representation of the word $w_i$. Let each vector representation $x_i$ be associated with a latent variable $z_i$, which denotes the topic or concept of a term and is an integer between 1 and K, where the parameter K is the total number of topics or the number of Gaussians in the mixture distribution. These latent variables, $z_i$s, can be estimated by an EM-based clustering algorithm such as K-means, where after the convergence of K-means on the set $V_d$, each $z_i$ represents the cluster id of each constituent vector $x_i$. Let the points $C_d = \{\mu_k\}_{k=1}^{K}$ represent the K cluster centres as obtained by the K-means algorithm. The posterior likelihood of the query to be sampled from the K Gaussian mixture model of a document $d^T$, centred around the $\mu_k$ centroids, can be estimated by the average distance of the observed query points from the centroids of the clusters, as shown in Equation (2).

$$P_{WVEC}(d^T|q^S) = \frac{1}{K|q|} \sum_i \sum_k \sum_j P(q_j^T|q_i^S) q_j^T \cdot \mu_k \tag{2}$$

In Equation (2), $q_j^T \cdot \mu_k$ denotes the inner product between the query word vector $q_j^T$ and the $k^{th}$ centroid vector $\mu_k$. Its weight is assigned with the values of $P(q_j^T|q_i^S)$ which denote the probability of translating a source word $q_i^S$ into the target-language word $q_j^T$. It is worth noting that a standard CLIR-based system is only capable of using the term overlap between the documents and the translated queries, and cannot employ the semantic distances between the terms to score the documents. In contrast, the set-based similarity, shown in Equation 2, is capable of using the semantic distances and therefore can be used to try to improve the performance of the alignment system.

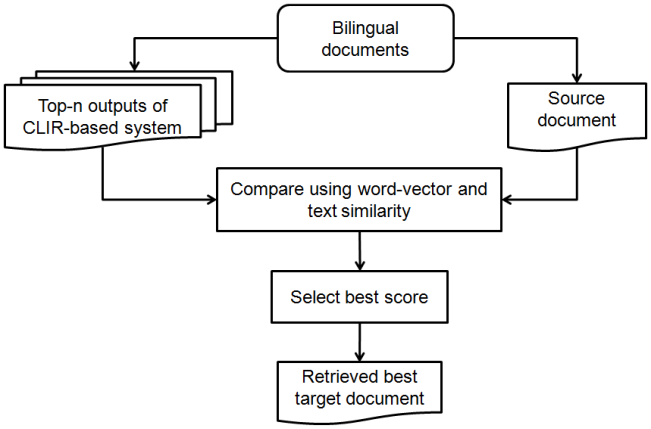


*Figure 2. Architecture of the word vector embedding-based system*

### 3.2.1. Combination with Text Similarity

Although the value of $P(d^T|q^S)$ is usually computed with the BoW representation model using language modeling (LM) (Ponte, 1998; Hiemstra, 2000) for CLIR (Berger and Lafferty, 1999), in our case we compute it with a different approach as shown in Equation (2). From a document $d^T$, the prior probability of generating a query $q^S$ is given by a multinomial sampling probability of obtaining a term $q_j^T$ from $d^T$. Then the term $q_j^T$ is transformed with the term $q_i^S$ in the source language. The priority belief (a parameter for LM) of this event is denoted by $\lambda$. As a complementary event to this, the term $q_j^T$ is also sampled from the collection and then transformed into $q_i^S$, with the prior belief $(1-\lambda)$. Let us consider that $P_{LM}(d^T|q^S)$ denotes this probability which is shown in Equation (3).

$$P_{LM}(d^T|q^S) = \prod_j \sum_i \lambda P(q_i^S|q_j^T)P(q_j^T|d^T) + (1-\lambda)P(q_i^S|q_j^T)P_{coll}(q_j^T) \quad (3)$$

In the next step, we introduce an indicator binary random variable to combine the individual contributions of the text-based and word vector-based similarity. Let us consider that this indicator is denoted by $\alpha$. We can then construct a mixture model of the two query likelihoods as shown in Equation (2) and Equation (3) for the word vector-based and the text-based methods, respectively. This combination is shown in Equation (4):

$$P(d^T|q^S) = \alpha P_{LM}(d^T|q^S) + (1-\alpha)P_{WVEC}(d^T|q^S) \quad (4)$$

#### 3.2.2. Construction of Index

The K-means clustering algorithm is run for the whole vocabulary of the words which can cluster the words into distinct semantic classes. These semantic classes are different from each other and each of them discusses a global topic (i.e., the cluster id of a term) of the whole collection. As a result of this, semantically related words are embedded in close proximity to each other.

While indexing each document, the cluster id of each constituent term is retrieved using a table look-up, so as to obtain the per-document topics from the global topic classes. The words of a document are stored in different groups based on their cluster-id values. Then the the cluster centroid of each cluster id is computed by calculating the average of the word vectors in that group. Consequently, we obtain a new representation of a document d as shown in Equation (5).

$$\mu_k = \frac{1}{|C_k|} \sum_{x \in C_k} x, C_k = \{x_i : c(w_i) = k\}, i = 1, \ldots, |d| \quad (5)$$

In the final step, the information about the cluster centroids is stored in the index. This helps to compute the average similarity between the query points and the centroid vectors during the retrieval process. The overall architecture of the word vector embedding-based approach is shown in Figure 2. It can be observed that this approach is combined with the text similarity method and makes use of the top-n outputs from the CLIR-based system to compare with the source document for which we intend to discover the alignment. In contrast, a system which is solely based on CLIR methodology simply re-ranks the top-n retrieved documents and selects the best one (as seen in Figure 1). Therefore, this extended version of our system facilitates the comparison of the document pair in terms of both the text and word-vector similarity as a continuation of our previous work (Lohar et al., 2016).

## 4. Experiments

### 4.1. Data

In all our experiments, we consider French as the source-language and English as the target language. Our experiments are conducted on two different sets of data, namely (i) *Euronews*[4] data extracted from the *Euronews* website[5] and (ii) the WMT '16[6] test dataset. The statistics of the English and French documents in the Euronews and the WMT-16 test datasets are shown in Table 1. The baseline system we use is based on

[4] https://github.com/gdebasis/cldocalign/tree/master/euronews-data

[5] http://www.euronews.com

[6] http://www.statmt.org/wmt16/

| dataset | English | French |
|---|---|---|
| Euronews | 40,419 | 39,662 |
| WMT-16 test dataset | 681,611 | 522,631 |

*Table 1. Statistics of the dataset.*

the Jaccard similarity coefficient[7] (JSC) to calculate the alignment scores between the document pair in comparison. This method focuses on the term overlap between the text pair and solves two purposes: (i) NE matches are extracted, and (ii) the common words are also taken into consideration.

In our initial experiments it was found that the Jaccard similarity alone produced better results than when combined with the cosine-similarity method or when only the cosine-similarity method was used. Therefore we decided to use only the former as the baseline system. We begin by using this method without employing any MT system and denote this baseline as 'JaccardSim'. Furthermore, we combine JaccardSim with the MT-output of the source-language documents to form our second baseline which is called 'JaccardSim-MT'.

### 4.2. Resource

The dictionary we use for the CLIR-based method is constructed using the EM algorithm in the IBM-4 word alignment (Brown et al., 1993) approach using the Giza++ toolkit (Och and Ney, 2003), which is trained on the English-French parallel dataset of Europarl corpus (Koehn, 2005). To translate the source language documents, we use Moses which we train on the English-French parallel data of Europarl corpus. We tuned our system on Euronews data and apply the optimal parameters on WMT test data.

## 5. Results

In the tuning phase, we compute the optimal values for the (empirically determined) parameters as follows; (i) $\lambda = 0.9$, (ii) $M = 7$, that is when we use 7 translation terms, and (iii) 60% of the terms from the document in order to construct the pseudo-query. The results on the Euronews data with the tuned parameters are shown in Table 2, where we can observe that the baseline approach (JaccardSim) has a quadratic time complexity (since all combinations of comparison are considered) and takes more than 8 hours to complete. In addition to this, the runtime exceeds 36 hours when combined with the MT system. In contrast, the CLIR-based approach takes only 5 minutes

[7] https://en.wikipedia.org/wiki/Jaccard_index

| Method | Parameters | | Evaluation Metrics | | | Run-time |
|---|---|---|---|---|---|---|
| | $\tau$ | M | Precision | Recall | F-score | (hh:mm) |
| JaccardSim | N/A | N/A | 0.0433 | 0.0466 | 0.0448 | 08:30 |
| JaccardSim-MT | N/A | N/A | 0.4677 | 0.5034 | 0.4848 | 36:20 |
| CLIR ($\lambda = 0.9$) | 0.6 | 7 | **0.5379** | **0.5789** | **0.5576** | **00:05** |

*Table 2. Results on the development set (EuroNews dataset).*

| Method | Parameters | | | | | Recall | Run-time |
|---|---|---|---|---|---|---|---|
| | $\lambda$ | $\tau$ | M | K | $\alpha$ | | (hhh:mm) |
| JaccardSim | N/A | N/A | N/A | N/A | N/A | 0.4950 | 130:00 |
| CLIR | 0.9 | 0.6 | 7 | N/A | N/A | 0.6586 | 007:35 |
| CLIR-WVEC | 0.9 | 0.6 | 7 | 20 | 0.9 | 0.6574 | 023:42 |
| CLIR-WVEC | 0.9 | 0.6 | 7 | 50 | 0.9 | **0.6619** | 024:18 |
| CLIR-WVEC | 0.9 | 0.6 | 7 | 100 | 0.9 | 0.6593 | 025:27 |

*Table 3. Results on the WMT test dataset.*

to produce the results. Moreover, the "JaccardSim" method has a very low effectiveness and can only lead to a considerable improvement when combined with MT. The CLIR-based approach produces the best results both in terms of precision and recall.

Table 3 shows the results on the WMT test dataset in which the official evaluation metric was only the recall measure to estimate the effectiveness of the document-alignment methods. However, we do not use "JacardSim-MT" system for the WMT dataset since it is impractical to translate a large collection of documents as it requires an unrealistically large amount of time.

We can draw the following observations from Table 3: (i) due to having a quadratic time complexity, the JaccardSim method has a high runtime of 130 hours. In contrast, the CLIR-based system is much faster and consumes only 7 hours. Additionally, it produces much higher recall than the JaccardSim method; (ii) the word-vector similarity method helps to further increase the recall produced by the CLIR-based approach, and (iii) a cluster value of 50 results in the highest value of recall among all values tested.

## 6. Conclusion and Future Work

In this paper we presented a new open-source multilingual document alignment tool based on a novel CLIR-based method. We proposed to use the measurement of the distances between the embedded word vectors in addition to using the term

overlap between the source and the target-language documents. For both the Euronews and WMT data, this approach produces a noticeable improvement over the Jaccard similarity-based baseline system. Moreover, an advantage of using the inverted index-based approach in CLIR is that it has a linear time complexity and can be efficiently applied to very large collections of documents. Most importantly, the performance is further enhanced by the application of the word vector embedding-based similarity measurements. We would like to apply our approach to other language pairs in future.

## Acknowledgements

This research is supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (`www.adaptcentre.ie`) at Dublin City University.

## Bibliography

Afli, Haithem, Loïc Barrault, and Holger Schwenk. Building and using multimodal comparable corpora for machine translation. *Natural Language Engineering*, 22(4):603 – 625, 2016.

Berger, Adam and John Lafferty. Information Retrieval As Statistical Translation. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 222–229, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. doi: 10.1145/312624.312681.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19: 263–311, June 1993. ISSN 0891-2017.

Hiemstra, Djoerd. *Using Language Models for Information Retrieval*. PhD thesis, Center of Telematics and Information Technology, AE Enschede, The Netherlands, 2000.

Koehn, Philipp. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, volume 5, pages 79–86, Phuket, Thailand, 2005.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Prague, Czech Republic, 2007.

Lohar, Pintu, Haithem Afli, Chao-Hong Liu, and Andy Way. The adapt bilingual document alignment system at wmt16. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, 2016.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of NIPS '13*, pages 3111–3119, Lake Tahoe, USA, 2013.

Munteanu, Dragos Stefan and Daniel Marcu. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504, 2005. ISSN 08912017.

Och, Franz Josef and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29:19–51, March 2003. ISSN 0891-2017.

Ponte, Jay Michael. *A language modeling approach to information retrieval*. PhD thesis, University of Massachusetts, MA, United States, 1998.

Resnik, Philip and Noah A. Smith. The Web as a parallel corpus. *Comput. Linguist.*, 29:349–380, September 2003. ISSN 0891-2017.

Roy, Dwaipayan, Debasis Ganguly, Mandar Mitra, and Gareth J. F. Jones. Representing Documents and Queries as Sets of Word Embedded Vectors for Information Retrieval. *CoRR*, abs/1606.07869, 2016.

Utiyama, Masao and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 72–79, Sapporo, Japan, 2003.

Yang, Christopher and Kar Wing Li. Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54:730–742, June 2003. ISSN 1532-2882. doi: 10.1002/asi.10261.

Zhao, Bing and Stephan Vogel. Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 745–748, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1754-4.

**Address for correspondence:**
Haithem Afli
haithem.afli@adaptcentre.ie
School of Computing, Dublin City University,
Dublin 9, Ireland