Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task

Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Emmanuel Dellandrea, Robert Gaizauskas, Mauricio Villegas and Krystian Mikolajczyk

Abstract. The ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task was the fourth edition of a challenge aimed at developing more scalable image annotation systems. In particular this year the focus of the three subtasks available to participants had the goal to develop techniques to allow computers to reliably describe images, localize the different concepts depicted in the images and generate a description of the scene. All three tasks use a single mixed modality data source of 500,000 web page items which included raw images, textual features obtained from the web pages on which the images appeared, as well as various visual features extracted from the images themselves. Unlike previous years the test set was also the training set and in this edition of the task hand-labelled data has been allowed. The images were obtained from the Web by querying popular image search engines. The development and subtasks 1 and 2 test sets were both taken from the "training set" and had 1,979 and 3,070 samples, and the subtask 3 track had 500 and 450 samples. The 251 concepts this year were chosen to be visual objects that are localizable and that are useful for generating textual descriptions of visual content of images and were mined from the texts of our large database of image-webpage pairs. This year 14 groups participated in the task, submitting a total of 122 runs across the 3 subtasks and 11 of the participants also submitted working notes papers. This result is very positive, in fact if compared to the 11 participants and 58 submitted runs of the last year it is possible to see how the interest in this topic is still very high.

1 Introduction

Every day, users struggle with the ever-increasing quantity of data available to them. Trying to find "that" photo they took on holiday last year, the image on Google of their favourite actress or band, or the images of the news article someone mentioned at work. There is a large number of images that can be cheaply found and gathered from the Internet. However, more valuable is mixed modality data, for example, web pages containing both images and text. A large amount of information about the image is present on these web pages and vice-versa. However, the relationship between the surrounding text and images varies greatly, with much of the text being redundant and/or unrelated. Moreover, images and the webpages on which they appear can be easily obtained for virtually any topic using a web crawler. In existing work such noisy data has indeed proven useful, e.g. [19,29,27]. Despite the obvious benefits of using such information in automatic learning, the very weak supervision it provides means that it remains a challenging problem. The Scalable Image Annotation, Localization and Sentence Generation task aims to develop techniques to allow computers to reliably describe images, localize the different concepts depicted in the images and generate a description of the scene.

The Scalable Image Annotation, Localization and Sentence Generation task is a continuation of the general image annotation and retrieval task that has been part of ImageCLEF since its very first edition in 2003. In the early years the focus was on retrieving relevant images from a web collection given (multilingual) queries, while from 2006 onwards annotation tasks were also held, initially aimed at object detection, but more recently also covering semantic concepts. In its current form, the 2015 Scalable Concept Image Annotation task is its fourth edition, having been organized in 2012 [23], 2013 [25] and 2014 [24]. In light of recent interest in annotating images beyond just concept labels, we introduced two new subtasks this year where participants developed systems to describe an image with a textual description of the visual content depicted in the image.

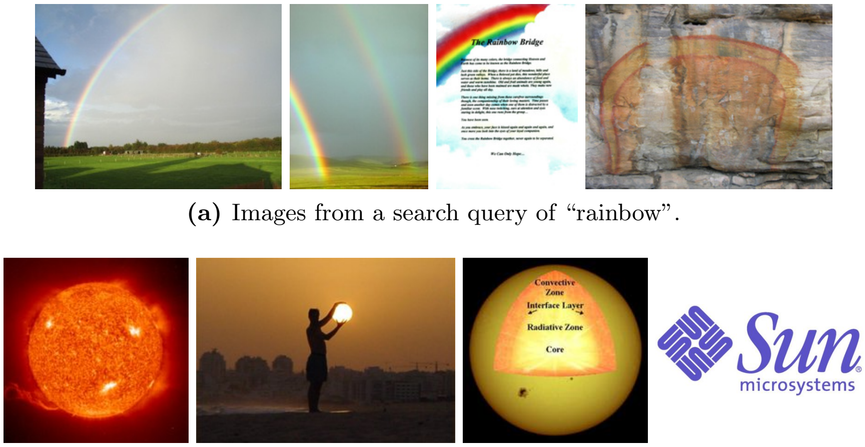
This paper presents the overview of the fourth edition of the Scalable Concept Image Annotation task [23,25,24], one of the four benchmark campaigns organized by ImageCLEF [21] in 2015 under the CLEF initiative[1]. Section 2 describes the task in detail, including the participation rules and the provided data and resources. Followed by this, Section 3 presents and discusses the results of the submissions. Finally, Section 4 concludes the paper with final remarks and future outlooks.

2 Overview of the Task

2.1 Motivation and Objectives

Image concept annotation, localization and natural sentence generation generally has relied on training data that has been manually, and thus reliably annotated, an expensive and laborious endeavour that cannot easily scale, particularly as the number of concepts grow. However, images for any topic can be cheaply gathered from the web, along with associated text from the webpages that contain the images. The degree of relationship between these web images and the surrounding text varies greatly, i.e., the data are very noisy, but overall these data contain useful information that can be exploited to develop annotation systems. Motivated by this need for exploiting this useful data, the ImageCLEF 2015 Scalable Concept annotation, localization and sentence generation task aims to develop techniques to allow computers to reliably describe images, localize the different concepts depicted in the images and generate a description of the scene. Figure 1 shows examples of typical images found by querying search engines. As can

[1] http://www.clef-initiative.eu



(a) Images from a search query of "rainbow".

(b) Images from a search query of "sun".

Fig. 1: Example of images retrieved by a commercial image search engine.

be seen, the data obtained are useful and furthermore a wider variety of images is expected, not only photographs but also drawings and computer generated graphics. This diversity has the advantage that this data can also handle the possible different senses that a word can have, or the different types of images that exist. Likewise, there are other resources available that can help to determine the relationships between text and semantic concepts, such as dictionaries or ontologies. There are also tools that can help to deal with noisy text commonly found on webpages, such as language models, stop word lists and spell checkers. The goal of this task was to evaluate different strategies to deal with the noisy data so that it can be reliably used for annotating, localizing, and generating natural sentences from practically any topic.

2.2 Challenge Description

This year the challenge[2] consisted of 3 subtasks

1. **Subtask 1:** The image annotation task continues in the same line of past years. The objective required the participants to develop a system that receives as input an image and produces as output a prediction of which concepts are present in that image, selected from a predefined list of concepts and starting this year, where they are located within the image.
2. **Subtask 2** (*Noisy Track*): In light of recent interest in annotating images beyond just concept labels, this subtask required the participants to describe images with a textual description of the visual content depicted in the image. It is thought of as an extension of subtask 1. This track was geared towards

[2] Challenge website at http://imageclef.org/2015/annotation

participants interested in developing systems that generated textual descriptions directly from images, e.g. by using visual detectors to identify concepts and generating textual descriptions from the detected concepts. This had a large overlap with subtask 1.

3. **Subtask 3** (*Clean track*): Aimed primarily at those interested only in the Natural Language Generation aspects of the subtask, therefore a gold standard input (bounding boxes labelled with concepts) was provided to develop systems that generate sentence, natural language based descriptions based on these gold standard annotations as input.

As common training set the participants were provided with 500,000 images crawled from the Internet, the corresponding webpages on which they appeared, as well as precomputed visual and textual features. Apart from the image and webpage data, the participants were also permitted and encouraged to use similar datasets and any other automatically obtainable resources to help in the processing and usage of the training data. In contrast to previous years, in this edition of the task hand labelled data has been allowed. Thus, the available trained ImageNet CNNs could be used, and the participants were encouraged to use also other resources such as ontologies, word disambiguators, language models, language detectors, spell checkers, and automatic translation systems. Unlike previous years, the test set was also the training set.

For the development of the annotation systems, the participants were provided with the following:

- A training dataset of images and corresponding webpages compiled specifically for the three subtasks, including precomputed visual and textual features (see Section 2.3).
- A development set of images with ground truth labelled bounding box annotations and precomputed visual features for estimating the system performance.
- A development set of images with at least five textual descriptions per image for **Subtask 2** and **Subtask 3**.
- A subset of the development set for **Subtask 3** with gold standard inputs (bounding boxes labelled with concepts) and correspondence annotation between bounding box inputs and terms in textual descriptions.

This year the training and the test images are all contained within the 500,000 images released at the beginning of the competition. At test time, it was expected that participants provided a classification for all images. After a period of two months, the development set, which included ground truth localized annotations, was released and about two months were given for participants to work on the development data. A maximum of 10 submissions per subtask (also referred to as *runs*) was allowed per participating group.

The 251 concepts this year were chosen to be visual objects that are localizable and that are useful for generating textual descriptions of the visual content of images. They include animate objects such as people, dogs and cats, inanimate objects such as houses, cars and balls, and scenes such as city, sea and mountains. The concepts were mined from the texts of our database of 31 million image-webpage pairs [22]. Nouns that are subjects or objects of sentences are extracted and mapped onto WordNet synsets [5]. These were then filtered to 'natural', basic-level categories (*dog* rather than a *Yorkshire terrier*), based on the WordNet hierarchy and heuristics from a large-scale text corpora [26]. The final list of concepts were manually shortlisted by the organisers such that they were (i) visually concrete and localizable; (ii) suitable for use in image descriptions; (iii) at a suitable 'every day' level of specificity that were neither too general nor too specific. The complete list of concepts, as well as the number of samples in the test sets, is included in Appendix A.

2.3 Dataset

The dataset[3] used was very similar to the one of the first three editions of the task [23,25,24]. To create the dataset, initially a database of over 31 million images was created by querying Google, Bing and Yahoo! using words from the Aspell English dictionary [22]. The images and corresponding webpages were downloaded, taking care to avoid data duplication. Then, a subset of 500,000 images was selected from this database by choosing the top images from a ranked list. For further details on how the dataset was created, please refer to [23]. The ranked list was generated by retrieving images from our database using the list of concepts, in essence, more or less as if the search engines had only been queried for these. From the ranked list, some types of problematic images were removed, and it was guaranteed that each image had at least one webpage in which they appeared.

The development and test sets were both taken from the "training set". A set of 5,520 images was selected for this purpose using a CNN trained to identify images suitable for sentence generation. The images were then annotated via crowd-sourcing in three stages: (i) image level annotation for the 251 concepts; (ii) bounding box annotation; (iii) textual description annotation. For the textual descriptions, basic spell correction was performed manually by the organisers using Aspell[4]. Both American and British English spelling variants (*color* vs. *colour*) were retained to reflect the challenge of real-world English spelling variants. A subset of these samples was then selected for subtask 3 and further annotated by the organisers with correspondence annotations between bounding box instances and terms in textual descriptions.

The development set contained 2,000 samples, out of which 500 samples were further annotated and used as the development set for the subtask 3. Note that only 1,979 samples from the development set contain at least one bounding box annotation. The number of textual descriptions for the development set ranged from 5 to 51 per image (with a mean of 9.5 and a median of 8 descriptions). The test set for subtasks 1 and 2 contains 3,070 samples, while the test set for subtask 3 comprises 450 samples which are disjoint from the test set of subtasks 1 and 2.

[3] Dataset available at http://risenet.prhlt.upv.es/webupv-datasets

[4] http://aspell.net/

Textual Data: Four sets of data were made available to the participants. The first one was the list of words used to find the image when querying the search engines, along with the rank position of the image in the respective query and search engine it was found on. The second set of textual data contained the image URLs as referenced in the webpages they appeared in. In many cases, the image URLs tend to be formed with words that relate to the content of the image, which is why they can also be useful as textual features. The third set of data was the webpages in which the images appeared, for which the only preprocessing was a conversion to valid XML just to make any subsequent processing simpler. The final set of data were features obtained from the text extracted near the position(s) of the image in each webpage it appeared in.

To extract the text near the image, after conversion to valid XML, the script and style elements were removed. The extracted texts were the webpage title, and all the terms closer than 600 in word distance to the image, not including the HTML tags and attributes. Then a weight $\mathrm{s}(t_n)$ was assigned to each of the words near the image, defined as

$$\mathrm{s}(t_n) = \frac{1}{\sum_{\forall t \in \mathcal{T}} \mathrm{s}(t)} \sum_{\forall t_{n,m} \in \mathcal{T}} F_{n,m} \, \mathrm{sigm}(\mathrm{d}_{n,m}) \ , \tag{1}$$

where $t_{n,m}$ are each of the appearances of the term t_n in the document $\mathcal{T}$, $F_{n,m}$ is a factor depending on the DOM (e.g. title, alt, etc.) similar to what is done in the work of La Cascia et al. [8], and $\mathrm{d}_{n,m}$ is the word distance from $t_{n,m}$ to the image. The sigmoid function was centered at 35, had a slope of 0.15 and minimum and maximum values of 1 and 10 respectively. The resulting features include for each image at most the 100 word-score pairs with the highest scores.

Visual Features: Before visual feature extraction, images were filtered and resized so that the width and height had at most 240 pixels while preserving the original aspect ratio. These raw resized images were provided to the participants but also eight types of precomputed visual features. The first feature set *Colorhist* consisted of 576-dimensional color histograms extracted using our own implementation. These features correspond to dividing the image in 3×3 regions and for each region obtaining a color histogram quantified to 6 bits. The second feature set *GETLF* contained 256-dimensional histogram based features. First, local color-histograms were extracted in a dense grid every 21 pixels for windows of size 41×41. Then, these local color-histograms were randomly projected to a binary space using 8 random vectors and considering the sign of the resulting projection to produce the bit. Thus, obtaining an 8-bit representation of each local color-histogram that can be considered as a *word*. Finally, the image is represented as a bag-of-words, leading to a 256-dimensional histogram representation. The third set of features consisted of *GIST* [13] descriptors. The following four feature types were obtained using the colorDescriptors software [17], namely *SIFT*, *C-SIFT*, *RGB-SIFT* and *OPPONENT-SIFT*. The configuration was dense sampling with default parameters and a hard assignment 1,000 codebook using a spatial pyramid of 1×1 and 2×2 [9]. Since the vectors of the spatial pyramid were concatenated, this resulted in 5,000-dimensional feature vectors. The codebooks were generated using 1.25 million randomly selected features and the k-means algorithm. And finally, *CNN* feature vectors have been provided computed as the seventh layer feature representations extracted from a deep CNN model pre-trained with the ImageNet dataset [15] using the Berkeley Caffe library[5].

2.4 Performance Measures

Subtask 1 Ultimately the goal of an image annotation system is to make decisions about which concepts to assign and localize to a given image from a predefined list of concepts. Thus to measure annotation performance, what should be considered is how good and accurate are those decisions the precision of a system. Ideally a recall measure would also be used to penalize a system that has additional false positive output. However given difficulties and unreliability of with the hand labeling of the concepts for the test images it wasn't possible to guarantee all concepts were labeled, however, it was assumed that the labels present were accurate and of a high quality.

The annotation and localization of Subtask 1 were evaluated using the PASCAL VOC [4] style metric of intersection over union (IoU), IoU is defined as

$$IoU = \frac{BB_{fg} \cap BB_{gt}}{BB_{fg} \cup BB_{gt}} \tag{2}$$

where BB is a rectangle bounding box, fg is a foreground proposed annotation label, gt is the ground truth label of the concept. It calculates the area of intersection between the foreground in the proposed output localization and the ground-truth bounding box localization, divided by the area of their union. IoU is superior to a more naive measure of the percentage of correctly labelled pixels as IoU is automatically normalized by the size of the object and penalizes segmentation's that include the background. This means that small changes in the percentage of correctly labelled pixels can correspond to large differences in IoU, and as the data-set has a wide variation in object size, the performance increases from our approach are more reliably measured. The evaluation of the ground truth and proposed output overlap was recorded from 0% to 90%. At 0%, this is equivalent to an image level annotation output, and 50% is the standard PASCAL VOC style metric used. The localized IoU is then used to compute the mean average precision (MAP) of each concept independently. This is then reported both per concept and averaged over all concepts.

Subtask 2 Subtask 2 was evaluated using the METEOR evaluation metric [2], which is an F-measure of word overlaps taking into account stemmed words, synonyms, and paraphrases, with a fragmentation penalty to penalize gaps and word order differences. This measure was chosen as it was shown to correlate

[5] More details can be found at https://github.com/BVLC/caffe/wiki/Model-Zoo

well with human judgments in evaluating image descriptions [3]. Please refer to Denkowski and Lavie [2] for details about this measure.

Subtask 3 Subtask 3 was also evaluated using the METEOR evaluation metric (see above). In addition, we have pioneered a fine-grained metric to evaluate the content selection capabilities of the sentence generation system. The *content selection* metric is the F_1 score averaged across all 450 test images, where each F_1 score is computed from the precision and recall averaged over all gold standard descriptions for the image. Intuitively, this measure evaluates how well the sentence generation system selects the correct concepts to be described against gold standard image descriptions. Formally, let $I = \{I_1, I_2, ...I_N\}$ be the set of test images. Let $G^{I_i} = \{G_1^{I_i}, G_2^{I_i}, ..., G_M^{I_i}\}$ be the set of gold standard descriptions for image I_i, where each $G_m^{I_i}$ represents the set of unique bounding box instances referenced in gold standard description m of image I_i. Let S^{I_i} be the set of unique bounding box instances referenced by the participant's generated sentence for image I_i. The precision P^{I_i} for test image I_i is computed as:

$$P^{I_i} = \frac{1}{M} \sum_m^M \frac{|G_m^{I_i} \cap S^{I_i}|}{|S^{I_i}|} \tag{3}$$

where $|G_m^{I_i} \cap S^{I_i}|$ is the number of unique bounding box instances referenced in both the gold standard description and the generated sentence, and M is the number of gold standard descriptions for image I_i.

Similarly, the recall R^{I_i} for test image I_i is computed as:

$$R^{I_i} = \frac{1}{M} \sum_m^M \frac{|G_m^{I_i} \cap S^{I_i}|}{|G_m^{I_i}|} \tag{4}$$

The content selection score for image I_i, F^{I_i}, is computed as the harmonic mean of P^{I_i} and R^{I_i}:

$$F^{I_i} = 2 \times \frac{P^{I_i} \times R^{I_i}}{P^{I_i} + R^{I_i}} \tag{5}$$

The final P, R and F scores are computed as the mean P, R and F scores across all test images.

The advantage of the macro-averaging process in equations (3) and (4) is that it implicitly captures the relative importance of the bounding box instances based on how frequently to which they are referred across the gold standard descriptions.

3 Evaluation Results

3.1 Participation

The participation was excellent, with a greater number of teams including a number of new groups. In total 14 groups took part in the task and submitted overall 122 system runs. The number of runs is nearly double the previous year. Among the 14 participating groups, 11 of them submitted a corresponding paper describing their system, thus for these there were specific details available. The following 14 teams submitted a working paper:

- **SMIVA** [7] The team from Social Media and Internet Vision Analytics Lab, the Institute for Infocomm Research, Singapore was represented by Pravin Kakar, Xiangyu Wang and Alex Yong-Sang Chia.
- **CEA LIST:** [6] The team from CEA, LIST, Laboratory of Vision and Content Engineering, France was represented by Etienne Gadeski, Herve Le Borgne, and Adrian Popescu.
- **CNRS TPT:** [16] The team from CNRS TELECOM ParisTech , France was represented by Hichem Sahbi.
- **RUC-Tencent:** [10] The team from RUC-Tenecent, 1Multimedia Computing Lab, School of Information, Renmin University of China was represented by Xirong Li, Qin Jin, Shuai Liao, Junwei Liang, Xixi He, Yu-Jia Huo, Weiyu Lan, Bin Xiao, Yanxiong Lu and Jieping Xu.
- **REGIM** [30] The team from REGIM: Research Groups on Intelligent Machines, University of Sfax, Tunisa was represented by Mohamed Zarka, Anis Ben Ammar and Adel Alimi.
- **Mindlab:** [14] The team from INAOE in Mexico and UNAL in Colombia was represented by Luis Pellegrin, Jorge A. Vanegas, John Arevalo, Viviana Beltrán, Hugo Jair Escalante, Manuel Montes-Y-Gómez and Fabio Gonzalez.
- **IVANLPR:** [11] The team from IVA Group, Chinese Academy of Sciences was represented by Yong Li, Jing Liu, Yuhang Wang, Bingyuan Liu, Jun Fu, Yunze Gao, Hui Wu, Hang Song, Peng Ying and Hanqing Lu..
- **KDEVIR:** [20] The team from Toyohashi University of Technology, Japan was represented by Md Zia Ullah and Masaki Aono..
- **UAIC:** [1] The team from UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania was represented by Alexandru Calfa and Adrian Iftene.
- **IRIP iCC:** [28] The team from Intelligent Recognition and Image Processing Lab, Beihang University, Beijing was represented by Yunhong Wang, Jiaxin Chen, Ningning Liu and Li Zhang.
- **LIP6:** [18] The team from Sorbonne Universits, CNRS, LIP6, Paris was represented by Ludovic Dos Santos, Benjamin Piwowarski and Ludovic Denoyer.

Table 6 provides the main key details for a number of the top groups submission describing their system. This table serves as a summary of the systems, and also is quite illustrative for quick comparisons. For a more in-depth look of the annotation systems of each team, please refer to their corresponding paper.

3.2 Results for Subtask 1

Subtask 1 was well received despite the additional requirement of labelling and localizing all 500,000 images. All submissions were able to provide results on all 500,000 images, indicating that all groups have developed systems that are scalable enough to annotate large amounts of images. The final results are presented in Table 1 in terms of mean average precision (MAP) over all images of all concepts, with both 0% overlap (i.e. no localization) and 50% overlap. It can be seen that three groups have achieved over 0.50 MAP across the evaluation set with 50% overlap with the ground-truth. This seems an excellent result given the challenging nature of the images used and the wide range of concepts provided. The graph in Figure 2 shows the performance of each submission for an increasing amount of overlap of the ground truth labels.

Group	**0% Overlap**	**50% Overlap**
SMIVA	0.79	0.66
IVANLPR	0.64	0.51
RUC	0.61	0.50
CEA	0.45	0.29
Kdevir	0.39	0.23
ISIA	0.25	0.17
CNRS-TPT	0.31	0.17
IRIP-iCC	0.61	0.12
UAIC	0.27	0.06
MLVISP6	0.06	0.02
REGIM	0.03	0.02
Lip6	0.04	0.01

Table 1: Subtask 1 results.

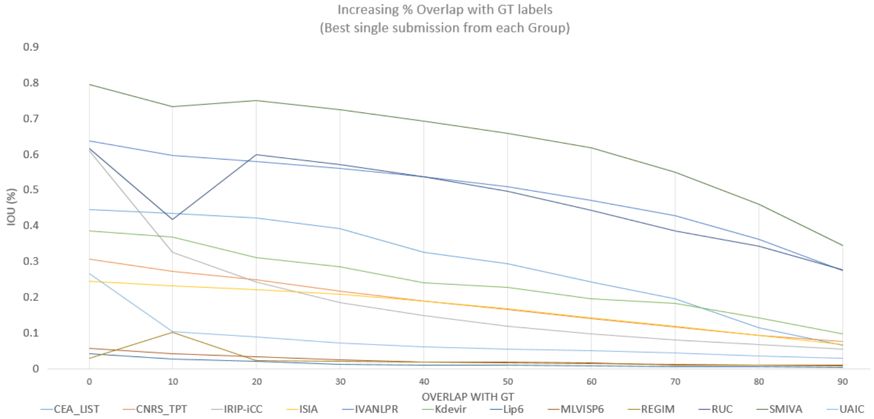


Fig. 2: Increasing Precision Overlap of submissions for sub task 1

The results from the groups seem encouraging and it would seem that the use of CNNs has allowed for large improvements in performance. Of the top 4 groups all use CNNs in their pipeline for the feature description.

SMIVA used a deep learning framework with additional annotated data, while **IVANLPR** implemented a two-stage process, initially classifying at an image level with an SVM classifier, and then applying deep learning feature classification to provide localization. While **RUC** trained a per concept, an ensemble of linear SVMs trained by Negative Bootstrap using CNN features as image representation. Concept localization was achieved by classifying object proposals generated by Selective Search. The approach by **CEA LIST** could be thought of as the baseline, they just use the CNN learnt features in a small grid based approach for localization.

Examples of the most and least successful localized concepts are shown in tables 2 and 3 respectively, together with the number of labelled occurrences of these concepts in the test data.

Concept	**Ave MAP across all Groups**	**Num. Occurrences**
Bee	0.64	5
Telephone	0.64	20
Fish	0.60	36
Suit	0.60	199
Mountain	0.60	77
Anchor	0.59	7
Bench	0.59	81
Fruit	0.58	17
Statue	0.58	84
Hog	0.54	24

Table 2: Successfully localized Concepts

Discussion for subtask 1 As can be observed in Table 1, the performance of many submissions was high this year, even given the additional constraint of localization. In fact, the 4 teams managed to achieve over 0.5 MAP, with 50% overlap with the ground truth. This perhaps indicates that in conjunction with the improvements from the CNN's real progress is starting to be made in the image annotation.

Figure 2 shows the change in performance as the requirements for intersection with the ground truth labels increases. All the approaches show a steady drop off in performance which is encouraging, illustrating that the approaches don't fail to detect a number of concepts correctly even with a high degree of accuracy. Even 90% overlap with the groundtruth the MAP for **SMIVA** was 0.35, which is impressive. Table 2 shows the most correctly localized concepts and also the

Concept	Ave MAP across all Groups	Num. Occurrences
Temple	0	26
Wheel	0	331
Letter	0	46
Apple	0	8
Cheese	0	1
Ribbon	0	45
Mushroom	0	6
leaf	0	134
rocket	0	9
Mattress	0	10

Table 3: Least Successfully localized Concepts

number of occurrences of the concept. As it is important to remember that due to imperfect annotation no recall level is calculated. This is likely to be why the concept *bee* is so high. However there is encouraging performance for *mountain*, *statue*, *bench* and *suit*. These are all quite varied concepts, in term of scale and percentage of the image the concept will cover. However examining Table 3 shows a number of concepts that should be detected and are not such as *leaf* and *wheel*. However, many in that table are quite small concepts and, therefore, harder to localize and intersect with the labelled ground truth. This could be an area to direct the challenge objectives in future years.

From a computer vision perspective, we would argue that the ImageCLEF challenge has two key differences in its dataset construction to that of the other popular data sets ImageNet [15] and MSCOCO [12]. All 3 are working on detection and classification of concepts within images. However, the ImageCLEF dataset is created from Internet web pages. This gives a key difference to the other popular datasets. The web pages are unsorted and unconstrained meaning the relationship or quality of the text and image in relation to a concept can be very variable. Therefore instead of a high-quality Flickr style photo of a car from ImageNet, the image in the ImageCLEF dataset could be a fuzzy abstract car shape in the corner of the image. This allows the ImageCLEF image annotation challenge to provide additional opportunities to test proposed approaches on. Another important difference is that in addition to the image, text data from web pages can be used to train and generate the output description of the image in a natural language form.

3.3 Results for Subtask 2

For subtask 2, participants were asked to generate sentence-level textual descriptions for all 500,000 training images. The systems were evaluated on a subset of 3,070 instances. Four teams participated in this pilot subtask. Table 4 shows the METEOR scores for subtask 3, for all submitted runs by all four participants.

Three teams achieved METEOR scores of over 0.10. **RUC** achieved the highest METEOR score, followed by **ISIA**, **MindLab**, and **UAIC**. We observed a large variety of approaches used by participants to tackle this subtask. **RUC** used the state of the art deep learning based CNN-LSTM caption generation system, **MindLab** employed a joint image-text retrieval approach, and **UAIC** a template-based approach.

As a comparison, we estimated a human upper-bound for this subtask by evaluating one description against the other descriptions for the same image and repeating the process for all descriptions. The METEOR score for the human upper-bound is estimated to be 0.3385 (Table 4). Therefore, there is clear scope for future improvement and work to improve image description generation systems.

3.4 Results for Subtask 3

For subtask 3, participants were provided with gold standard labelled bounding box inputs for 450 test images (released one week before the submission deadline), and were asked to generate textual descriptions for each image based on the gold standard input bounding boxes. To enable evaluation using the content selection metric (Section 2.4), participants were also asked to indicate within the textual descriptions the bounding box(es) to which the relevant term(s) correspond.

Two teams participated in this subtask (both of whom also participated in subtask 2). Table 5 shows the content section and METEOR scores for the subtask, again for all submitted runs by the two participants. **RUC** performed marginally better than **UAIC** in terms of the F and METEOR scores. Interestingly, it can be observed that **RUC**'s sentence generation system has higher precision P, while **UAIC** achieved higher recall R in general than **RUC**. This is possibly due to **RUC**'s use of a deep learning based sentence generator coupled with re-ranking based on the gold standard input which yielded higher precision, while **UAIC**'s template-based generator selected more bounding boxes to be described resulting in a higher recall. Note that the METEOR scores are generally higher in subtask 3 compared to subtask 2 as participants are provided with gold standard input concepts, as well as the subtask having a smaller test set of 450 samples.

As a baseline, we generated textual descriptions per image by selecting at most three bounding boxes from the gold standard at random (the average number of unique instance mentions per description in the development set is 2.89). These concepts terms were then connected with random words or phrases selected randomly from a predefined list of prepositions and conjunctions followed by an optional article *the*. Like subtask 2, we also computed a human upper-bound. The results for these are shown in Table 5. As observed, all participants performed significantly better than the random baseline. Compared to the human upper-bound, again much work can still be done. An interesting note is that **RUC** achieved a high precision P almost on par with the human upper-bound, at the expense of a lower R.

Group	**Run**	**METEOR**			
		Mean ± Std	**Median**	**Min**	**Max**
RUC	1	0.1659 ± 0.0834	0.1521	0.0178	0.5737
	2	0.1781 ± 0.0853	0.1634	0.0178	0.5737
	3	0.1806 ± 0.0817	0.1683	0.0192	0.5696
	4	0.1759 ± 0.0860	0.1606	0.0178	0.5737
	5	0.1684 ± 0.0828	0.1565	0.0178	0.5696
	6	0.1875 ± 0.0831	0.1744	0.0201	0.5696
ISIA	1	0.1425 ± 0.0796	0.1269	0.0151	0.5423
	2	0.1449 ± 0.0811	0.1295	0.0000	0.5737
	3	0.1644 ± 0.0842	0.1495	0.0181	1.0000
	4	0.1502 ± 0.0812	0.1352	0.0000	0.5737
	5	0.1687 ± 0.0852	0.1529	0.0387	1.0000
MindLab	1	0.1255 ± 0.0650	0.1140	0.0194	0.5679
	2	0.1143 ± 0.0552	0.1029	0.0175	0.4231
	3	0.1403 ± 0.0564	0.1342	0.0256	0.3745
	4	0.1230 ± 0.0531	0.1147	0.0220	0.5256
	5	0.1192 ± 0.0521	0.1105	0.0000	0.4206
	6	0.1260 ± 0.0580	0.1172	0.0000	0.4063
	7	0.1098 ± 0.0527	0.1005	0.0000	0.4185
	8	0.1079 ± 0.0498	0.1004	0.0000	0.3840
	9	0.0732 ± 0.0424	0.0700	0.0135	0.2569
	10	0.1202 ± 0.0528	0.1123	0.0000	0.5256
UAIC	1	0.0409 ± 0.0310	0.0309	0.0142	0.2954
	2	0.0389 ± 0.0286	0.0309	0.0142	0.2423
	3	0.0483 ± 0.0389	0.0331	0.0142	0.2954
	4	0.0813 ± 0.0513	0.0769	0.0142	0.3234
Human	-	0.3385 ± 0.1556	0.3355	0.0000	1.0000

Table 4: Results for subtask 2, showing the METEOR scores for all runs from all participants. We consider the mean METEOR score as the primary measure, but for completeness we also present the median, min and max scores.

Group	Run	Content Selection Score			METEOR
		Mean *F*	**Mean *P***	**Mean *R***	
RUC	1	0.5310 ± 0.2327	0.6845 ± 0.2999	0.4771 ± 0.2412	0.2393 ± 0.0865
	2	0.5147 ± 0.2390	0.7015 ± 0.3095	0.4496 ± 0.2488	0.2213 ± 0.0845
UAIC	1	0.4201 ± 0.1938	0.4582 ± 0.2410	0.4481 ± 0.2467	0.1709 ± 0.0771
	2	0.4701 ± 0.1678	0.4520 ± 0.1743	0.5447 ± 0.2398	0.2055 ± 0.0589
	3	0.5021 ± 0.1774	0.5130 ± 0.1939	0.5496 ± 0.2409	0.2080 ± 0.0654
	4	0.5023 ± 0.1774	0.5093 ± 0.1934	0.5534 ± 0.2420	0.2093 ± 0.0661
	5	0.5030 ± 0.1775	0.5095 ± 0.1938	0.5547 ± 0.2415	0.2097 ± 0.0660
Baseline	-	0.1800 ± 0.1973	0.1983 ± 0.2003	0.1817 ± 0.2227	0.0977 ± 0.0467
Human	-	0.7445 ± 0.1174	0.7690 ± 0.1090	0.7690 ± 0.1090	0.4786 ± 0.1706

Table 5: Results for subtask 3, showing the content selection and the METEOR scores for all runs from all participants.

3.5 Limitations of the challenge

There are two major limitations that we have identified with the challenge this year. Very few of the groups used the provided data set and features, we found this surprising, considering the state of the art CNN features and many others were included. However, this is likely to be due to the complexity and challenge of the 500,000 web page based images. Given they were collected from the Internet with little, a large number of the images are poor representations of the concept. In fact a number of participants annotated a large amount of their own more perfect training data, as their learning process assumes perfect or near perfect training examples, it will fail. As the number of classes increases and become more varied annotating all perfect data will become more difficult.

Another shortcoming of the overall challenge is the difficulty of ensuring the ground truth has 100% of concepts labelled, thus allowing a recall measure to be used. This is especially problematic as the concepts selected include fine-grained categories such as *eyes* and *hands* that are generally small but occur frequently in the dataset. In addition, it was difficult for annotators to reach a consensus in annotating bounding boxes for less well-defined categories such as *trees* and *field*. Given the current crowd-source based hand-labelling of the ground truth, the concepts have missed annotations. Thus, in this edition a recall measure is not evaluated for subtask 1.

4 Conclusions

This paper presented an overview of the ImageCLEF 2015 Scalable Concept Image Annotation task, the fourth edition of a challenge aimed at developing more scalable image annotation systems. The focus of the three subtasks available to participants had the goal to develop techniques to allow computers to reliably annotate images, localize the different concepts depicted in the images and generate a description of the scene.

The participation increased this year compared to last year with 14 teams submitting in total 122 system runs. The performance of the submitted systems was somewhat superior to last year's results for sub task 1. Especially considering the requirement to label all 500,000 images in the training/test set. This was in part probably due to the increased CNN usage as the feature representation. The clear winner of this year's subtask 1 evaluation was the SMIVA [7] team, which placed heavy emphasis on the visual aspect of annotating images and improved their overall annotation performance by branching off secondary recognition pipelines for certain highly common concepts. The participation rate for subtasks 2 and 3 is encouraging as pilot subtasks. For subtask 3, we also pioneered a concept selection metric to encourage fine-grained evaluation of image descriptions. RUC [10] led both subtasks using the state of the art CNN-LSTM caption generator, improving performance by exploiting concept detections from subtask 1. Other teams, however, varied in their approaches to the problem. The encouraging participation rate and promising results in these pilot subtasks are sufficient motivations for them to be included in future editions of the challenge.

The results of the task have been very interesting and show that useful annotation systems can be built using noisy web crawled data. Since the problem requires to cover many fronts, there is still a lot of work that can be done, so it would be interesting to continue this line of research. Papers on this topic should be published, demonstration systems based on these ideas be built and more evaluation of this sort be organized. Also, it remains to see how this can be used to complement systems that are based on clean hand labeled data and find ways to take advantage of both the supervised and unsupervised data.

Table 6: Key details of the best system for top performing groups that submitted a paper describing their system.

System	Visual Features [Total Dim.]	Other Used Resources	Training Data Processing Highlights	Annotation Technique Highlights
SMIVA [7]	1024-dim GoogLeNet [1] [T.Dim. = 21312]	* WordNet * Bing Image Search	Training data created by augmenting target concept with WordNet hyponyms and lemmas, retrieving images from Bing Image Search and filtering out too small or uniform images.	Uses selective search to generate object proposals, runs classifiers on each proposal and performs non-maximal suppression. Secondary pipelines add further context/processing from faces and difficult to localize concepts (e.g. trees).
IVANLPR [11]	ImageNet CNNs	-	Annotation by classification with deep visual features and linear SVM. Annotation by search with surrounding text.	Localization by Fast RCNN for concepts with obvious object. Localization by search for the scene related concepts.
RUC-Tencent [10]	Caffe CNNs	* Flickr Images	Hierarchical Semantic Embedding (HierSE) for selecting positive examples, Negative Bootstrap for building concept classifiers.	Selective Search for generating object proposals and refinement to reduce false alarms.
CEA LIST [6]	ImageNet CNNs [T.Dim. = 256]	* Bing Image Search	The network is trained with noisy web data corresponding to the concepts to detect in this task - just using simple CNNs.	Cell based regions to localize the concepts.

Acknowledgments

The Scalable Image Annotation, Localization and Sentence Generation task was co-organized by the ViSen consortium under the EU CHIST-ERA D2K Programme, supported by EPSRC Grants EP/K01904X/1 and EP/K019082/1, and by French ANR Grant ANR-12-CHRI-0002-04. This work was also supported by the European Science Foundation (ESF) through the research networking programme Evaluating Information Access Systems (ELIAS).

References

1. Calfa, A., Iftene, A.: Using Textual and Visual Processing in Scalable Concept Image Annotation Challenge. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)
2. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation (2014)
3. Elliott, D., Keller, F.: Comparing automatic evaluation measures for image description. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 452–457. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/P14-2074
4. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision 111(1), 98–136 (Jan 2015)
5. Fellbaum, C. (ed.): WordNet An Electronic Lexical Database. The MIT Press, Cambridge, MA; London (May 1998)
6. Gadeski, E., Borgne, H.L., Popescu, A.: CEA LIST's participation to the Scalable Concept Image Annotation task of ImageCLEF 2015. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)
7. Kakar, P., Wang, X., Chia, A.Y.S.: SMIVA at ImageCLEF 2015: Automatic Image Annotation using Weakly Labelled Web Data. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)
8. La Cascia, M., Sethi, S., Sclaroff, S.: Combining textual and visual cues for content-based image retrieval on the World Wide Web. In: Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on. pp. 24–28 (1998), doi:10.1109/IVL.1998.694480
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2. pp. 2169–2178. CVPR '06, IEEE Computer Society, Washington, DC, USA (2006), doi:10.1109/CVPR.2006.68
10. Li, X., Jin, Q., Liao, S., Liang, J., He, X., Huo, Y.J., Lan, W., Xiao, B., Lu, Y., Xu, J.: RUC-Tencent at ImageCLEF 2015: Concept Detection, Localization and Sentence Generation. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)
11. Li, Y., Liu, J., Wang, Y., Bingyuan Liu, J.F., Gao, Y., Wu, H., Song, H., Ying, P., Lu, H.: Hybrid Learning Framework for Large-Scale Web Image Annotation and Localization. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)

12. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR abs/1405.0312 (2014), http://arxiv.org/abs/1405.0312
13. Oliva, A., Torralba, A.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. Int. J. Comput. Vision 42(3), 145–175 (May 2001), doi:10.1023/A:1011139631724
14. Pellegrin, L., Vanegas, J.A., Arevalo, J., Beltrán, V., Escalante, H.J., Montes-Y-Gómez, M., Gonzalez, F.: INAOE-UNAL at ImageCLEF 2015: Scalable Concept Image Annotation. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) pp. 1–42 (April 2015)
16. Sahbi, H.: CNRS TELECOM ParisTech at ImageCLEF 2015 Scalable Concept Image Annotation Task: Concept Detection with Blind Localization Proposals. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)
17. van de Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating Color Descriptors for Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1582–1596 (2010), doi:10.1109/TPAMI.2009.154
18. Santos, L.D., Piwowarski, B., Denoyer, L.: Graph Based Method Approach to the ImageCLEF2015 Task1 - Image Annotation. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)
19. Torralba, A., Fergus, R., Freeman, W.: 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 30(11), 1958–1970 (nov 2008), doi:10.1109/TPAMI.2008.128
20. Ullah, M.Z., Aono, M.: KDEVIR at ImageCLEF 2015 Scalable Image Annotation, Localization, and Sentence Generation task: Ontology based Multi-label Image Annotation. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)
21. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A.G.S., Bromuri, S., Amin, M.A., Mohammed, M.K., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., del Mar Roldán García, M.: General Overview of ImageCLEF at the CLEF 2015 Labs. Lecture Notes in Computer Science, Springer International Publishing (2015)
22. Villegas, M., Paredes, R.: Image-Text Dataset Generation for Image Annotation and Retrieval. In: Berlanga, R., Rosso, P. (eds.) II Congreso Español de Recuperación de Información, CERI 2012. pp. 115–120. Universidad Politécnica de Valencia, Valencia, Spain (June 18-19 2012)
23. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop, Online Working Notes. Rome, Italy (September 17-20 2012), http://mvillegas.info/pub/Villegas12_CLEF_Annotation-Overview.pdf
24. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In: CLEF2014 Working Notes. CEUR Workshop Proceedings, vol. 1180, pp. 308–328. CEUR-WS.org, Sheffield, UK (September 15-18 2014), http://ceur-ws.org/Vol-1180/CLEF2014wn-Image-VillegasEt2014.pdf

25. Villegas, M., Paredes, R., Thomee, B.: Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013), http://mvillegas.info/pub/Villegas13_CLEF_Annotation-Overview.pdf
26. Wang, J.K., Yan, F., Aker, A., Gaizauskas, R.: A poodle or a dog? Evaluating automatic image annotation using human descriptions at different levels of granularity. In: Proceedings of the Third Workshop on Vision and Language. pp. 38–45. Dublin City University and the Association for Computational Linguistics, Dublin, Ireland (August 2014), http://www.aclweb.org/anthology/W14-5406
27. Wang, X.J., Zhang, L., Liu, M., Li, Y., Ma, W.Y.: ARISTA - image search to annotation on billions of web photos. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 2987–2994 (June 2010), doi:10.1109/CVPR.2010.5540046
28. Wang, Y., Chen, J., Liu, N., Zhang, L.: BUAA-iCC at ImageCLEF 2015 Scalable Concept Image Annotation Challenge. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)
29. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: learning to rank with joint word-image embeddings. Machine Learning 81, 21–35 (2010), doi:10.1007/s10994-010-5198-3
30. Zarka, M., Ammar, A.B., Alimi., A.: Regimvid at ImageCLEF 2015 Scalable Concept Image Annotation Task: Ontology based Hierarchical Image Annotation. In: CLEF 2015 Evaluation Labs and Workshop, Online Working Notes (2015)

A Concept List 2015

The following tables present the 251 concepts used in the ImageCLEF 2015 Scalable Concept Image Annotation task. In the electronic version of this document, each concept name is a hyperlink to the corresponding WordNet synset webpage.

Concept	WordNet		#dev.	#test
	type	sense#		
airplane	noun	1	22	76
anchor	noun	1	-	7
apple	noun	1	3	8
apron	noun	1	2	28
arm	noun	1	83	4352
bag	noun	1	37	150
bag	noun	4	28	88
ball	noun	1	36	63
balloon	noun	1	7	12
banana	noun	1	2	2
barn	noun	1	6	4
baseball_glove	noun	1	10	27
basin	noun	1	2	20
basket	noun	1	12	7
bat	noun	1	-	-
bathroom	noun	1	5	8
bathtub	noun	1	2	1
beach	noun	1	27	5
bear	noun	1	7	20
beard	noun	1	22	178
bed	noun	1	32	31
bee	noun	1	1	5
beer	noun	1	3	10
bell	noun	1	1	-
bench	noun	1	36	81
bicycle	noun	1	30	56
bin	noun	1	22	49
bird	noun	1	14	48
blackberry	noun	1	-	1
blanket	noun	1	17	55
boat	noun	1	76	104
bomb	noun	1	1	5
book	noun	1	30	45
boot	noun	1	19	101
bottle	noun	1	42	81
bouquet	noun	1	-	-
bowl	noun	1	12	24
box	noun	1	28	86
bread	noun	1	8	6
brick	noun	1	21	116
bridge	noun	1	34	80
bucket	noun	1	9	19
bullet	noun	1	2	2
bus	noun	1	25	94
butter	noun	1	2	-
butterfly	noun	1	1	1
cabinet	noun	1	29	89
camera	noun	1	18	37
can	noun	1	8	4
canal	noun	1	5	13
candle	noun	1	7	9
candy	noun	1	2	30
cannon	noun	1	4	13
cap	noun	1	67	223
car	noun	1	181	603
cat	noun	1	5	20
cathedral	noun	1	15	58
cave	noun	1	4	5
ceiling	noun	1	21	124
chair	noun	1	111	448

Concept	WordNet		#dev.	#test
	type	sense#		
cheese	noun	1	1	1
city	noun	1	37	36
cliff	noun	1	9	22
clock	noun	1	5	3
computer	noun	1	14	41
corn	noun	1	-	-
cow	noun	1	19	66
crab	noun	1	3	3
cross	noun	1	4	30
cup	noun	1	20	96
curtain	noun	1	41	127
dam	noun	1	2	2
deer	noun	1	13	57
dish	noun	1	13	71
dog	noun	1	49	76
doll	noun	1	8	11
door	noun	1	87	429
dress	noun	1	100	384
drill	noun	1	3	-
drum	noun	1	13	25
dryer	noun	1	2	-
ear	noun	1	27	1803
egg	noun	1	-	1
elephant	noun	1	9	23
eye	noun	1	39	2783
face	noun	1	43	3205
fan	noun	1	4	2
farm	noun	1	3	3
feather	noun	1	2	3
female_child	noun	1	72	206
fence	noun	1	94	423
field	noun	1	185	163
fireplace	noun	1	9	8
fish	noun	1	9	36
flag	noun	1	35	131
flashlight	noun	1	1	2
floor	noun	1	69	327
flower	noun	1	96	359
foot	noun	1	14	1291
fork	noun	1	7	5
fountain	noun	1	10	7
fox	noun	1	-	5
frog	noun	1	1	2
fruit	noun	1	6	17
garden	noun	1	35	142
gate	noun	1	12	58
goat	noun	1	12	7
grape	noun	1	-	7
guitar	noun	1	26	42
gun	noun	1	20	34
hair	noun	1	121	2644
hallway	noun	1	13	82
hammer	noun	1	3	2
hand	noun	1	170	3455
hat	noun	1	92	391
head	noun	1	30	3861
helicopter	noun	1	8	16
helmet	noun	1	51	186
hill	noun	1	19	85

continues in next page

Concept	WordNet		#dev.	#test
	type	sense#		
hog	noun	3	1	24
hole	noun	1	1	6
hook	noun	1	1	11
horse	noun	1	58	83
hospital	noun	1	1	2
house	noun	1	135	725
jacket	noun	1	60	654
jean	noun	1	51	370
key	noun	1	1	5
keyboard	noun	1	10	6
kitchen	noun	1	9	8
knife	noun	1	5	8
ladder	noun	1	14	32
lake	noun	1	28	74
leaf	noun	1	116	134
leg	noun	1	30	3185
letter	noun	1	13	46
library	noun	1	2	1
lighter	noun	2	1	537
lion	noun	1	9	5
lotion	noun	1	-	4
magazine	noun	1	7	20
male_child	noun	1	89	260
man	noun	1	681	2962
mask	noun	1	12	15
mat	noun	1	6	5
mattress	noun	1	3	10
microphone	noun	1	27	67
milk	noun	1	1	1
mirror	noun	1	19	75
monkey	noun	1	4	7
motorcycle	noun	1	22	61
mountain	noun	1	85	77
mouse	noun	1	1	1
mouth	noun	1	48	1568
mushroom	noun	1	-	6
neck	noun	1	14	1400
necklace	noun	1	50	37
necktie	noun	1	33	210
nest	noun	1	1	2
newspaper	noun	1	16	26
nose	noun	1	16	1970
nut	noun	1	1	2
office	noun	1	9	3
onion	noun	1	-	-
orange	noun	1	1	9
oven	noun	1	1	6
painting	noun	1	45	156
pan	noun	1	2	4
park	noun	1	27	344
pen	noun	1	11	14
pencil	noun	1	4	5
piano	noun	1	9	9
picture	noun	1	25	158
pillow	noun	1	19	48
planet	noun	1	-	1
pool	noun	1	23	20
pot	noun	1	4	17
potato	noun	1	3	2
prison	noun	1	-	-
pumpkin	noun	1	1	9
rabbit	noun	1	5	11
rack	noun	1	10	1
radio	noun	1	1	14
ramp	noun	1	3	3
ribbon	noun	1	11	45
rice	noun	1	-	-
river	noun	1	51	82
rock	noun	1	94	239
rocket	noun	1	4	9
rod	noun	1	7	31
rug	noun	1	35	52
salad	noun	1	1	2
sandwich	noun	1	3	5
scarf	noun	1	23	67
sea	noun	1	107	215
sheep	noun	1	7	10
ship	noun	1	50	183
shirt	noun	1	153	1946
shoe	noun	1	59	1145
shore	noun	1	41	93
short_pants	noun	1	39	368
signboard	noun	1	91	624
skirt	noun	1	16	120
snake	noun	1	9	6
sock	noun	1	7	185
sofa	noun	1	36	62
spear	noun	1	1	-
spider	noun	1	1	-
stadium	noun	1	27	99
star	noun	1	2	1
statue	noun	1	35	84
stick	noun	1	17	156
strawberry	noun	1	-	1
street	noun	1	143	440
suit	noun	1	77	199
sunglasses	noun	1	45	144
sweater	noun	1	33	107
sword	noun	1	5	5
table	noun	2	125	320
tank	noun	1	7	10
telephone	noun	1	6	20
telescope	noun	1	4	1
television	noun	1	10	29
temple	noun	1	14	26
tent	noun	1	10	57
theater	noun	1	2	19
toilet	noun	1	5	5
tongue	noun	1	4	17
towel	noun	1	6	20
tower	noun	1	32	93
town	noun	1	10	199
tractor	noun	1	7	7
train	noun	1	13	27
tray	noun	1	3	28
tree	noun	1	460	1444
truck	noun	1	44	86
tunnel	noun	1	3	3
valley	noun	1	13	29
vase	noun	1	14	26
vest	noun	1	10	113
wagon	noun	1	6	14
wall	noun	1	104	855
watch	noun	1	29	93
waterfall	noun	1	1	4
well	noun	1	-	1
wheel	noun	1	52	331
wicket	noun	1	-	5
window	noun	1	134	1308
wine	noun	1	10	25
wolf	noun	1	2	1
woman	noun	1	474	1491