The JeuxDeMots Project is 10 Years Old: What We have Learned

Alain Joubert, Mathieu Lafourcade, Nathalie Le Brun
LIRMM, Montpellier, France - Imaginat, Lunel, France
alain.joubert@lirmm.fr, mathieu.lafourcade@lirmm.fr, imaginat@imaginat.name

Abstract

Here we present the assessment of 10 years of experience concerning the JDM project, a set of GWAPs for NLP, among which a main game combined with many satellite games aims to build a large lexical-semantic network for the French language. We highlight the lessons learned from this experience for creating lexical resources through a never ending process. We emphasize that combining automatic inference processes with player activity is particularly relevant to build such a resource. **Keywords:** crowdsourcing, game with a purpose, inferences, lexical semantic network

Introduction

The JeuxDeMots (JDM) project, whose very first GWAP was launched 10 years ago, in July 2007 (Lafourcade, 2007), aims to build a very large lexical-semantic network for the French language. Such a resource is usable in any application needing some semantic analysis of textual information and some reasoning capabilities about world fact and common sense. As a graph, the lexical network contains terms (words, groups of words, expressions, inflected forms, and symbolic informations) connected by typed semantic relations. It was an ambitious project, in the same spirit as Wordnet (Miller, 1995) for the aimed goal, and experience showed us it was feasible: the resource is freely available (C0 license) with a monthly updated export. Building such a resource may be made through different ways:

manual acquisition is a costly, long and fastidious work, where information would not likely be updated without further funding (the typical example being Wordnet or Framenet (Baker et al., 1998));

automatic construction from corpora: result can be biased by the corpus itself or the extraction method. Moreover, to correctly extract semantic relations, it is necessary to carry out a semantic analysis, which is precisely the object of the resource that one wishes to build;

myriadization of paid parcel work, with the risk that the data obtained are not of the expected quality (Fort et al., 2011); this type of method is based on the fact that many Internet contributors, often referred to as *turkers*, are willing to collaborate and are generally (lowly) remunerated for this collaboration.

Hence, we developed a collaborative game on the web in a crowd-sourcing way, where players would not-knowingly construct the resource by playing. As far as we know, prior 2007, such a method had never been used in the NLP domain.

In this paper, we first recall on which principles the main game relies, before addressing the adjustments we have had to perform. Then we present the generated resource, and the automatic methods that densify the network by consolidating (correction and completion) the data obtained from the games. We also point out several useful aspects of such a network in the field of NLP. Then, we discuss the lessons learned after ten years of using the JDM model. Finally, we emphasize that combining automatic inference processes with player activity is particularly relevant to build and densify such a resource.

1. JeuxDeMots

JDM is a GWAP (Game With A Purpose, see Von Ahn (VonAhn, 2004) and (Lafourcade et al., 2015)), that is to say a collaborative game which has a definite purpose beside entertaining (for example, collecting data or solving problems).

In a JDM match, two players collaborate anonymously and in an asynchronous way. A match of JDM is to propose a term and an instruction, asynchronously, to two players who do not know each other, and then to confront their answers. For example : *Give generics of goldfish*, or *which are the parts of motor-vehicle ?* Each player has a limited time to provide the answers he / she deems relevant, and when both sets of answers are confronted, the system only retains the common answers, to limit the risk of error: it is believed that answer is likely to be relevant when given by two players who have had no opportunity to consult each other. When both players give the same answer, but it does not exist in our database, the term is added. At the end of a game both players are rewarded with points and virtual gifts.

The number of terms and relations increases through player activity: we started with a $150,000$ terms data base and no relation; ten years later (2017), the network has more than 2.6 million terms and 180 million relations.

As we said, JDM is a game, but it is a useful game. The play aspect is essential to attract and retain players, and make them going on participating. But it's also a useful game, and as designers of JDM, we must never lose sight that the goal of the game is to build a lexical network. We will return in detail on this dual aspect in section 3, when we will develop what 10 years of JDM experience has taught us.

1.1. Evolutions of the Game

For a game to be attractive and attractive, to avoid monotony is essential. That's why we have tried to develop different game modes, to stimulate the emulation between the players by all sorts of rankings; we created the possibility for the players to give themselves gift-parties, to challenge themselves to duels, to choose from about 30 "skills" (ie the type of relations on which to answer, as synonym,

cause, consequence, family, agent, patient, instrument, location, feature, part, etc.) and to test many other parameters of play. The idea was to offer the possibility to play the main game in all sorts of ways, with all kinds of configuration. Moreover, we have gradually created, in addition to the main game, 12 "satellite" games, so that a player can temporarily abend the main game to try another game, and thus participate in the consolidation and verification of the data obtained through the main game. Indeed the analysis of the first data made some adjustments of the main game necessary, but also gave us the idea to create new games to verify, reinforce, or correct some data.

For the main game of JDM, the turn-over is relatively high: most players are active for about 3 weeks, sometimes even for several months, even years... Some have been playing JDM for 10 years! The initial game has therefore benefited from many improvements and additions over time, as it is detailed in (Lafourcade et al., 2015). We will highlight in particular:

The opportunity to **retry your chance** after a disappointing game, and even to sue the other player. The trials are held in public, the other players play the role of jurors, which is yet another way to create animation and conviviality.

The ability to **play on the theme of his choice**: a player will give more relevant answers in a field in which he is expert or passionate.

The ability to **choose the level of difficulty** of the proposed terms. It is strategic to offer some easy vocabulary (e.g. *tiger* or *land*) to a novice player, so that he is not discouraged and earns points quickly. But after a few games, most players prefer harder terms (e.g. *Higgs boson*), and it's adjustable in the game's options.

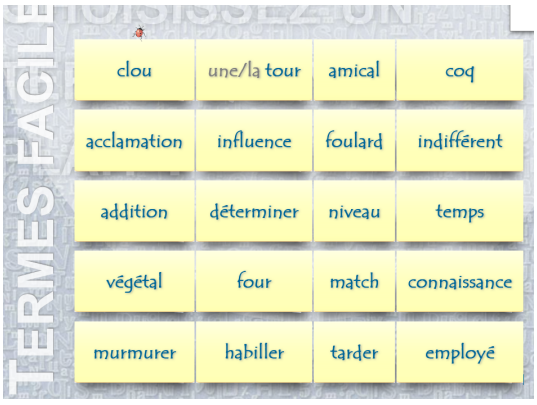

Figure 1: Selecting easy terms in JeuxDeMots amongst a randomized set of terms.

The ability to **offer games**, with the terms and relations of their choice, to other players (and even to attach to this gift a personal message). It is a way of entrusting the sampling of terms to the players, and thus increase their productivity: the players spontaneously choose interesting term / relation pairs, that is to say for which there are many and interesting answers.

A **chat** to communicate in real time with other connected players or the JDM administrator. This reinforces the sense of belonging to a community and allows them to help each other, to help newcomers and to guide them in discovering the many features of the games, to explain how to answer for difficult relations, exchange "tricks" to play better and earn more points, etc.

Figure 2: Offered Gifts in JeuxDeMots, allowing a relevant sampling of terms to be annotated.

However, the most important development was the creation of these "satellite" games, in addition to the main game. For players, these games offer another type of interaction: many are click or vote games, fast, easily playable on a smarphone in common situations such as in a waiting room or public transport. For the lexical network under construction, these "satellite" games compensate for the bias of the main game. Some of them, like Totaki, validate the data collected (Joubert et al., 2011), others, like Askit, correct errors related to polysemy, others focus on specific types of relations: polarity of terms for likeIt, feelings and emotions for Emot, colors and appearance for ColorIt... Tierxical helps refine the relations weighting, Askyou allows to validate or invalidate pending proposals, etc.

1.2. Evolution of Players

It soon turned out that a significant percentage of players, very interested in the "purpose" dimension of the GWAPs, expressed the desire to take a more active and concrete part in the construction of the lexical network. It is for these players that was set up the Diko, a contributive interface: the volunteer players can go and make contributions directly in the entry and for the relation(s) that inspire them. (Lafourcade et al., 2015). This role of active contributor is well suited to people sensitive to the challenge of participatory or citizen science.

To minimize the risk of error related to these contributors, who remain amateurs, a system of validation of their contributions by majority vote has been set.

2. Obtained Resource: a Very Large Knowledge Base

The lexical-semantic network (dubbed RezoJDM), under permanent construction, has been produced by the players, contributors and automated inference mechanisms (aka bots) and can be considered as a knowledge base encompassing both common sense, specialized and lexical informations.

Figure 3: A given play of JeuxDeMots and its outcome.

In addition to being typed (for example: *r_isa*, *r_agent*, *r_patient*, *r_domain*, etc.) a relation is also weighted. Its weight depends on the number of players who have proposed it. A weight can be negative (< 0), it indicates a negative relation (for example, *an ostrich **can not** fly*). Similarly, a false relation is made negative rather than deleted, to keep in mind that it was proposed and then invalidated. Thus, since relations can be proposed by automated processes, negatively weighting a false relation avoids the system to propose the same erroneous relation in a recurring way. Thus, inference mechanisms can also rely on negative relations.

Needless to say that this resource evolves over time with the addition of new terms and relations (at the very least new named entities). Its construction is not supposed to ended one day (at least theoretically).

Since the startup, the network gained on average around 20000 terms and 1.4 million semantic relations each month. Although the progression is not strictly regular, it is globally linear in time and we do not observe (yet) a beginning of flattening of the progression curve.

2.1. Common Sense & Domain Knowledges

The RezoJDM is a knowledge base containing mostly common sens facts. In order to process texts from specific domains, some efforts were done to integrate specialty domains, for example in health domain (anatomy, medicine, radiology, oncology) (Lafourcade and Ramadier, 2016) or in culinary domain (cooking, ingredients, nutritional facts) (Clairet and Lafourcade, 2017).

2.2. Densification with Automatic Inference

New relations can be inferred from existing ones through automatic endogenous inference, or from other (external resources) by extracting exogenous semantic relations.

Endogenous inferences rely on mechanisms of deduction, induction and abduction (Zarrouk and Lafourcade, 2014). For example : *a cat is a feline* and *a feline has part claws*, so we can deduce that probably *a cat has part claws*.

Exogenous extraction of semantic relations is undertaken from other resources, such as Wikipedia (Lafourcade and Joubert, 2013), or from fictions (French literature) corpora or non fiction and journalistic (Le Monde) corpora.

The contributions are tagged with the name of their author, whether human or automatic mechanism and are pending validation, either through satellite games, or by a game administrator. As shown in (Zarrouk and Lafourcade, 2014), inferred relationships may be wrong, especially when the inference is made from polysemous terms. Manual intervention by an expert is then required.

2.3. Error Detection

Even though the error rate is relatively low in the JDM network, well below 0.1%, we have developed an automatic error detection mechanism, (Lafourcade et al., 2017). which, from a so-called "primary" error, reported by a player or a contributor, will detect and report the errors secondarily induced by the automatic mechanisms of inference.

3. Lessons from the JDM Experience

Our 10 years of experience and exploitation of the JDM model have allowed us to identify a number of characteristics that a GWAP must have in order to be sustainable. (Lafourcade and Joubert, 2013).

3.1. About the Gameplay

Ideally, a GWAP should:

- be attractive, fun and interesting, which is essential to attract a large number of players: such a game must present a ludic interest at the interface level to attact gamers, but even more at the content level in order to keep them;
- be easy to understand, both in terms of the game modes and instructions to respect; a too complex game, or requiring a long learning, will discourage a large number of players;
- arouse addiction : this is possible thanks to the features of the game, as for example the instant replay by simple click, but also the modalities of play and the possibilities of interaction with the other players (lawsuits, gifts, theft of words, duels,...) that encourage people to come back;
- allow the filtering of players : flatter and make them feel useful (which is true) but also make them feel guilty if they do not play well (eventually make them give up the game if they do not improve). It's a good way to keep only the good players and guarantee the quality of the produced resource.

3.2. Benefits for NLP

The durability of a GWAP certainly depends on its attractiveness to the players, but also on how it meets the expectations of its designer. He must be able, by comparing the data he gets to what he wanted to obtain, to make the adjustments and modifications necessary to obtain usable data. The advantages for the NLP community are multiple:

- The data obtained is the result of non-negotiated contributions since the two players whose answers will be confronted have no way of communicating.
- The resource obtained is low cost compared to that which would be built manually, and it is acquired quickly (more than 40000 relations per day);
- The data acquisition procedure is ethical, unlike other approaches, such as Amazon Mechanical Turk (Fort et al., 2011). The principle of GWAP does not raise any ethical problem as long as it remains free and does not offer prizes that look like disguised salaries.

3.3. Issues in Cheating and Vandalism

As shown in (Lafourcade et al., 2015), we also noticed some cases of cheating and vandalism:

- Cheating : some players have managed to bypass some restrictive game rules, such as time limitation. This kind of cheating does not question the quality of the resource obtained, but it may disgust and discourage the players who do not cheat, and this can result in a disaffection for the game. In the context of JDM, we noticed that it was the first hours which constituted the critical phase for this type of risk.
- Vandalism is intended to corrupt the database by knowingly inserting erroneous data. Designers must minimize this risk at all costs, as detecting errors introduced is quite difficult, and must be done manually by experts. In fact, we think it is almost impossible to detect this type of error in an automatic manner. The fact that we only validate the common answers of a pair of players who do not know each other limits the risk of vandalism. As a result, assuming that the system could be able to detect an incongruous information (which is already far from being obvious and which poses the insolvable question of criteria), to systematically classify it as wrong and eliminate it would be counterproductive: incongruous does not necessarily mean wrong.

4. Impact of Automated Inferences

As mentioned above, automatic extraction or inference of semantic relation is at the core of the development of the lexical network.

4.1. Bots Behaving as Players

We recall the principle of the game: the game of a player is compared to another game on the same term and the same instruction (type of relation), and the common answers supply the network. The other part is randomly selected by the system. How to be sure that a game with the same term and the same instruction is available?

To deal with this issue, we devised fake player (bot) which produce pending games when needed. Of course, for a given term and, if they are enough true player games available, no bot is invocated for generating games. The state of the network directly dictates the nature and quality of the bot's answers. In such a way, along with players, the network feeds itself.

Player bots make use of various strategies, but the principle is to select proposals (randomly between 10 and 40) from the network according to three criteria: a) the most activate relations, b) the least activated relations, and c) the relations waiting to be (in)validated. Thus, player bot may induce the validation of waiting and original contributions.

One should notice that a bot never plays against itself but only against true players. A player bot never contributes directly to the network, it does only indirectly through games done with human players.

The average number of common responses between a bot player and a human player is about 12, while that number is about 5 between two human players.

4.2. Bots Behaving as Contributors

Thanks to automatic mechanisms of inference, robots act as contributors and add relations to the network. These relations are proposals, which must be validated by the human players-contributors, who vote for or against. As mentioned above, the inference is done according to different approaches: deduction, induction, and various types of abduction.

Moreover, some bots are able to deduce certain rules from the structure of the network. A rule is a) a set of conditions that must be verified for a given term and b) a conclusion wich is a relation to be added to the term. For example : *$x r_isa 'animal aquatique'* $\rightarrow$ *$x r_lieu 'eau'* (Eng: *if $x is a kind of aquatic animal then $x could be located in water*). A bot proposes a rule as soon as it finds at least 3 examples and no counter-example (negative relations). If validated (by human administrator), the rule is applied to the network and the found conclusion is directly inserted (no validation required). So far, 4469 rules have been validated and led to the automatic creation of over 50 million relations (out of 180 million in January 2018).

So far, the error rate of automatic contributions is less than 1 for 10000 and 97% of such errors have been automatically detected.

4.3. Snowball Effect

The automatic inference mechanisms work from what is already validated in the network. To give a simple example of inference based on deduction, if we know that *pigeon* is_a *oiseau* (*bird*), then *pigeon* will inherit the general properties of *bird* (that is to say, semantic relations of *bird* with other words).

As a mean, each relation introduced by a player in the lexical network leads to 57 new correct relations inferred (from various bots and strategies), and the number of incorrect proposed relations tends to decrease as the network grows, from 20 in 2012 to 13 in 2014 and finally 5 in 2016.

We estimated that without the action of bot-players nor the mechanisms of inference the number of relations in RezoJDM would be of around 3 million, instead of more than 180. In addition, as the snowball effect results in an increase in the number of relationships validated automatically via the game (because the number of common responses between a bot and a human player is statistically higher than between two human players), both quality and quantity of the data collected is much better than it would have been with only human players.

5. Conclusion

The JDM project has largely demonstrated the interest of combining GWAPs and inference mechanisms to build a reliable and large-scale lexico-semantic resource. More precisely, this resource has been built largely by the activity of players and direct contributors, but also critically supplemented by mechanisms of automatic inferences. Those mechanisms have been instrumental concerning the significant volume and quality of the resource.

Our approach is monolingual and language independent. As a research perspective, we are currently developing a multilingual game similar to JDM, with which we expect to obtain a very large lexical database in a large amount of various languages. Such an approach could be especially instrumental in collecting cross-lingual lexical information for languages with a reduced number of speakers.

References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bos, J. and Nissim, M. (2015). Uncovering noun-noun compound relations by gamification. In Beáta Megyesi, editor, *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 251–255.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase Detectives: a web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*.

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013). Using games to create language resources: Successes and limitations of the approach. *Gurevych I., Kim J. (eds) The People's Web Meets NLP. Theory and Applications of Natural Language Processing. Springer*, oct.

Clairet, N. and Lafourcade, M. (2017). Towards the automatic detection of nutritional incompatibilities based on recipe titles. In *CD-MAKE 2017*, pages 346–366. Reggio di Calabria, Italy, Aug 29-Sep 1, 2017.

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Joubert, A., Lafourcade, M., Schwab, D., and Zock, M. (2011). Evaluation et consolidation d'un réseau lexical via un outil pour retrouver le mot sur le bout de la langue. In *Proc. of the 18th conference on Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier.

Lafourcade, M. and Joubert, A. (2013). Bénéfices et limites de l'acquisition lexicale dans l'expérience jeuxdemots. In *Ressources Lexicales: Contenu, construction, utilisation, évaluation*, pages 187–216. Linguisticae Investigationes, Supplementa 30, John Benjamins.

Lafourcade, M. and Ramadier, L. (2016). Semantic relation extraction with semantic patterns experiment on radiology reports. In *10th edition of the Language Resources and Evaluation Conference (LREC 2016)7*. Portoroz, Slovenia, Aug 23-28 May 2016, 6 p.

Lafourcade, M., Joubert, A., and Le Brun, N. (2015). Games with a purpose (gwaps). page 158. Wiley-ISTE, July.

Lafourcade, M., Joubert, A., and Le Brun, N. (2017). If mice were reptiles, then reptiles could be mammals or how to detect errors in the jeuxdemots lexical network? In *Proc.of International Conference on Recent Advances on Natural Language Processing (RANLP 2017)*, Varna, Bulgaria, September.

Lafourcade, M. (2007). Making people play for lexical acquisition. In *Proc. of the 7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand, December.

Liu, H. and Singh, P. (2004). Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, oct.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, December.

VonAhn, L. (2004). Labelling images with a computer game. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 319–326.

Zarrouk, M. and Lafourcade, M. (2014). Inferring knowledge with word refinements in a crowdsourced lexical-semantic network. In *In proc. of the the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 346–366. Dublin, Irlande, 2014, 9 p.